Een communicatief lexicon door het verankeren van woordbetekenissen. Oratie bij de aanvaarding van het ambt van bijzonder hoogleraar Computationale Lexicologie bij de Faculteit der Letteren van de Vrije Universiteit Amsterdam

Vossen, P.T.J.M.

2006

document version
Publisher's PDF, also known as Version of record

Een communicatief lexicon door het verankeren van woordbetekenissen

‘You don’t know what love is ’til you’ve learned the meaning of the blues’
(Raye/DePaul, 1941)

prof.dr. P.Th.J.M. Vossen

Rede uitgesproken in verkorte vorm bij de aanvaarding van het ambt van bijzonder hoogleraar Computationele Lexicologie vanwege de Stichting Het Vrije Universiteitsfonds bij de faculteit der Letteren van de Vrije Universiteit Amsterdam op 22 december 2006.

vrije Universiteit *amsterdam*

Opgedragen aan mijn vader Jan Vossen (†) en mijn schoonvader Jan Weisscher.

"You don't know what love is 'til you've learned the meaning of the blues."
(Ray/DePaul, 1941)

1 Inleiding

In dit citaat, staat "the blues" voor het diepste voelen en realiseren, en staat 'love' voor een woord dat abstract blijft totdat je voelt wat je mist. De tekst gaat verder met "Until you've loved a love you've had to lose, you don't know what love is".

Deze zin verankert de betekenis van 'liefde' aan het ondergaan van een gevoel en het afwezig zijn van dat gevoel, het missen. De ware betekenis van het woord wordt hiermee gelijkgesteld aan een persoonlijke en diepe ervaring. De interessante vraag is dan, in hoeverre is die betekenis overdraagbaar als die teruggaat op een persoonlijke diepe ervaring? In hoeverre kunnen we de betekenis van zo'n woord communiceren? Hoe kun je aan een kind uitleggen wat de betekenis van liefde is? Hoe wordt ***liefde*** dan beschreven, uitgelegd?

Meestal gebeurt dat heel indirect, bijna als de diagnose van een ongrijpbare ziekte, die we dienen te herkennen:

- hartkloppingen, rood worden, droge mond, kortademig
- wegdromen, niet-geconcentreerd, onverschillig voor wereldse zaken
- geen eetlust, veel luisteren naar muziek, poëzie
- vlinders in je buik
- enz.

Het is dus kennelijk niet zo makkelijk om de betekenis van dat woord aan iemand duidelijk te maken en daarom wellicht onderwerp van zoveel teksten en kunst. Nu is het woord 'liefde' wel een heel moeilijk voorbeeld om de betekenis van woorden te bespreken, maar het illustreert wel de essentie van een filosofische en wetenschappelijke discussie over wat betekenis is.

De algemene consensus is namelijk dat men gelooft dat het principieel onmogelijk is om de betekenis van taal, in al zijn aspecten, formeel te definiëren en vast te leggen. Als je erover nadenkt is dat een bizarre conclusie. Hoe is het mogelijk dat mensen communiceren zonder een formele en bewijsbare definitie van betekenis? Hoe is het mogelijk dat we beslissingen nemen, kennis overdragen, beschavingen bouwen zonder dat er sprake is van echt begrip en echte betekenis? Is er dan sprake van een 'beetje' betekenis, kun je elkaar tot op zekere hoogte begrijpen? Als dat zo is, tot waar gaat die hoogte en hoe komen we erachter wat wel en wat geen overdraagbare betekenis heeft?

Vooruitgang in het beschrijven van mentale en neurologische processen zou misschien uitkomst kunnen bieden. Onze hersenen worden meer en meer in kaart gebracht en de modellen van de fysische en ook functionele werking zullen in de nabije toekomst steeds verfijnder worden. Maar zelfs als we er ooit in zullen slagen om gevoelens en ervaringen als neurologische processen te kunnen meten, vastleggen (filmen of registreren) of zelfs modelleren in computersimulaties, zelfs dan kunnen we de ervaringsfilm van de ene persoon niet vergelijken met de ervaringsfilm van een andere persoon.

In de praktijk wordt betekenis dan ook alleen beperkt vastgelegd, hetzij doordat alleen die betekenis of dat deel van de betekenis wordt gedefinieerd dat die zich laat vangen in een

logische formule, hetzij doordat betekenisafspraken alleen indirect of globaal worden gemaakt. In het eerste geval, wordt betekenis vaak vereenvoudigd tot een wiskundig model van variabelen en relaties. In een dergelijk model is 'liefde' de relatie ***L*** tussen ***x*** en ***Y***: ***L***(x,y): er is sprake van *liefde* dan en alleen dan indien er een wereld bestaat met een ***x*** en een ***y*** waartussen de relatie ***L*** waar is.

In het tweede geval, worden woorden vervangen door andere woorden zoals in een woordenboek, in de hoop dat die duidelijk(er) maken wat er bedoeld wordt: *warme genegenheid, gehechtheid aan een persoon of zaak* (Groot Woordenboek der Nederlandse Taal, 1992). Iemand die de definities van woordenboeken volgt door de betekenis van de woorden in de definitie ook weer op te zoeken houdt uiteindelijk niet veel over:

- genegenheid = welwillende gezindheid jegens iemand, in sterkere opvatting naderend tot liefde
- gezindheid = innerlijke houding
- houding = (fig.) wijze van handelen en optreden, gedrag, manier van handelen

(bron: Groot Woordenboek der Nederlandse Taal, 1992)

Door het vervangen van woorden door woorden wordt niets fundamenteels verklaard en komen we alleen verder af te staan van wat we eigenlijk bedoelen.

Het moge duidelijk zijn dat het beide benaderingen aan veel 'blues' ontbreekt. Wat is er dan veranderd in de wereld dat deze patstelling zou kunnen doorbreken? Er zijn eigenlijk twee ontwikkelingen die een nieuwe situatie hebben gecreëerd. De eerste ontwikkeling is dat in de huidige maatschappij de meeste communicatie digitaal is geworden en dus voor het eerst in de geschiedenis van de mensheid op grote schaal meetbaar. Digitale communicatie en digitale informatie opgeslagen in tekstuele vorm in vele talen. Een simpele search met Google naar het woord liefde of het equivalent in andere talen levert miljoenen voorkomens op:

Resultaten **1 - 10** van circa **1.180.000** voor **liefde** (**0,05** seconden)
Resultaten **1 - 10** van circa **114.000.000** voor **liebe** (**0,09** seconden)
Resultaten **1 - 10** van circa **924.000.000** voor **love** (**0,04** seconden)
Resultaten **1 - 10** van circa **81.800.000** voor **amour** (**0,06** seconden)
Resultaten **1 - 10** van circa **115.000.000** voor **amor** (**0,07** seconden)
Resultaten **1 - 10** van circa **358.000.000** voor **愛**(**0,25** seconden) , Kanji, Japans (ai)

(bron: http://www.google.nl, 11 november 2006)

Naast allerlei andere zaken is er dus kennelijk veel *liefde* op het Internet te vinden. Het woord wordt hier dus in allerlei contexten gebruikt en we kunnen het vinden in de omgeving van allerlei andere woorden. Dit biedt unieke mogelijkheden om niet alleen ons gedrag in kaart te brengen in relatie tot een woord (wat voor websites bezigen dit woord?) maar ook om het woord *liefde* in verband te brengen met andere woorden, bijvoorbeeld ***relaties***. Er is dus een schat aan data en materiaal om de betekenis van woorden te bestuderen.

Maar er is meer aan de hand. Iedere dag zoeken miljoenen mensen naar informatie en ze doen dat meestal aan de hand van vragen in natuurlijke taal. Het proces van het zoeken naar informatie door het stellen van vragen in taal is een communicatief proces dat kan worden opgeslagen, worden opgemeten op een ongekende schaal. In al die communicatieve processen speelt de betekenis van woorden een grote rol. Het is mogelijk om die betekenis op verschillende manieren vast te leggen en zelfs te manipuleren en op die manier het communicatieve effect te beïnvloeden. Mensen gebruiken taal om aan een systeem vragen te

stellen en die over te brengen aan andere gebruikers van taal. Een eenvoudig voorbeeld is een *dating service*, waar mensen ***liefde*** zoeken (en soms vinden) door de betekenis van woorden in taal.

Sterker nog, het is mogelijk dat computersystemen actief deelnemen aan die communicatieve processen. Doordat betekenis meer en meer wordt vastgelegd in informatie, zullen computersystemen in staat zijn om informatie steeds beter met elkaar te kunnen verbinden. Kennis zal daardoor steeds sneller groeien en daardoor ook de dienstbaarheid van systemen aan mensen. Mensen zullen moeten communiceren met die computersystemen om behoeften kenbaar te maken en die systemen zullen die behoefte moeten begrijpen en om kunnen zetten in een nuttig resultaat. In de toekomst zou een computerprogramma zelfs op zoek kunnen gaan naar jouw ideale liefde!

Wat er veranderd is, is dat communicatieprocessen zijn verworden tot een systeem dat kan worden gemonitord maar dat ook zelf actief kan deelnemen. Niet alleen informatie en taal is digitaal beschikbaar maar communicatieve modellen gaan een steeds grotere rol spelen in de manier waarop die informatie kan worden gebruikt. Betekenis van woorden speelt hierin een cruciale rol. Het heeft dus wel degelijk ***zin*** om de betekenis van woorden zo goed mogelijk te ***verankeren*** in een systeem dat daar nuttig mee om kan gaan. Dit ondanks de wellicht fundamentele bezwaren die er zijn tegen het definiëren van betekenis. Met ***verankeren*** bedoel ik dat het communicatieve effect van het woord nuttig of effectief is in allerlei communicatieve situaties, tussen computers en mensen, tussen computers en gegevens en tussen mensen onderling met bemiddeling of via een computerondersteund systeem.

Het begrip betekenis verwordt zo dus tot een ‘engineering probleem’: hoe effectieve communicatie en dus dienstbaarheid tot stand te brengen door het vastleggen van betekenis in een informatie- en kennismaatschappij? Woorden krijgen dan betekenis voor een mens en voor een computer als een instructietaal die wel of niet effectief is in het bereiken van een doel. Weliswaar geformuleerd als een engineering probleem, is dat dan misschien uiteindelijk ook de essentie van betekenis, ook tussen mensen onderling.

In de moderne communicatie en informatiemaatschappij is een nieuwe situatie ontstaan waarin het op allerlei manieren bepalen en vastleggen van wat woorden betekenen een dagelijks terugkerend probleem is, dat nationale, talige en culturele grenzen overschrijdt. Een database met de woorden van een taal heeft niets meer weg van een woordenboek als naslagwerk voor eigen gebruik. Het heeft eerder het karakter van een vastgelegde standaard of zelfs computercode die bepaalt hoe communicatieve processen tussen mensen en machines of tussen machines onderling gaan verlopen. Op termijn, wellicht gaan bepalen hoe communicatie tussen mensen gaat verlopen. Een uitgever is gewaarschuwd: verander niet zomaar de betekenis van een woord, de maatschappij zou kunnen ontwrichten.

In deze rede wil ik een kader voorstellen waarin betekenis van woorden en uitdrukkingen bestudeerd kan worden èn kan worden vastgelegd oftewel ***verankerd***, vanuit het perspectief van een wereld waarin informatie, kennis en communicatie meer en meer wordt gemodelleerd en gestructureerd door computersystemen. Het proces van verankeren zie ik daarin als langzaam en arbeidsintensief, waarin de speelruimte van woorden meer en meer wordt ingeperkt. De betekenis van woorden kan altijd worden opgerekt en aangepast, maar woorden kunnen ook niet zomaar alles betekenen. Een woord is niet een boot die eeuwig op drift is en zomaar overal kan opduiken. Er zijn veel manieren om een boot te ***verankeren*** (aan de bodem, aan een kade of zelfs aan elkaar) en dat geldt ook voor de woorden in een taal: aan

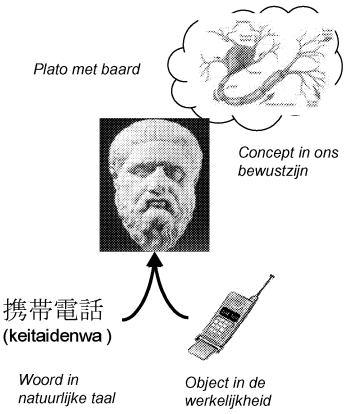
andere woorden, aan woorden in andere talen en aan andere feiten in de werkelijkheid. Juist door het probleem van de betekenis van woorden vanuit verschillende disciplines aan te pakken is het mogelijk om dichter bij de waarheid of inhoud te komen. Daarom spelen in dit kader verschillende disciplines een rol:

1. De werkelijkheid van personen en dingen in de wereld zoals die meer en meer wordt vastgelegd in standaarden en dataobjecten;
2. De eigenschappen van talen die bepalen wat hoe wordt uitgedrukt en wat kan worden uitgedrukt. Wat is de relatie van taal met een gestandaardiseerde definitie van de werkelijkheid?
3. Meertaligheid en multiculturele samenlevingen waarin communicatie meer en meer plaatsvindt. In hoeverre kan communicatie worden beregeld ongeacht taal- en cultuurverschillen?
4. Modellen die beschrijven hoe mensen zich gedragen in communicatieve processen;
5. Internet en informatietechnologie die bepaalt hoe informatie wordt opgeslagen en toegankelijk wordt gemaakt;

Al deze aspecten spelen een belangrijke rol in het verankeren en dus nuttig maken van betekenis van woorden. In de volgende paragrafen zal ik deze invalshoeken een voor een toelichten.

2 Betekenis als filosofisch probleem

De betekenis van 'betekenis' is van oudsher een filosofische kwestie. Het westerse denken over betekenis, is gegrondvest in het denken van twee klassieke grootheden: Plato en Aristoteles. Het gaat daarbij sinds de oudheid om de relatie tussen een ***woord***, ***iets*** in de ***werkelijkheid*** en wat er in ons hoofd afspeelt (iets ***mentaals***). Deze relatie is door Odgen en Richards (1923) schematisch weergegeven in een driehoek:



Figuur 1: Driehoeksrelatie tussen woorden, concepten en werkelijkheid, gebaseerd op Odgen en Richards.

In dit plaatje staat het Japanse ***woord*** 携帯電話 (*keitaidenwa, keitai*) symbool voor een ***concept*** in het hoofd van Plato. Dit concept is hier weergegeven als een configuratie van neuronen oftewel een individuele gedachte. Het concept verwijst naar een ***object*** in de werkelijkheid dat we als zodanig waarnemen, in dit geval een *mobiele telefoon*.

Als we het plaatje van de mobiele telefoon wegdenken, dan wordt onmiddellijk duidelijk dat wij, als niet-sprekers van het Japans, geen idee hebben wat er in het hoofd van Plato omgaat bij het horen van het woord 携帯電話 (*keitaidenwa*). Het woord als klank en als geschreven tekst is betekenisloos zonder een gedeelde werkelijkheid.

Volgens Quine (1964) blijft het woord echter altijd betekenisloos, óók als we het observeren in samenhang met een mobiele telefoon in de werkelijkheid. Wij als niet-sprekers van het Japans weten nooit zeker dat 携帯電話 (*keitaidenwa*) ook het juiste woord is voor mobiele telefoon, sterker nog, zelfs Japanners hebben geen absolute zekerheid. Zoals in de inleiding gesteld is, is het onmogelijk om te bepalen wat er in ons omgaat bij het zien van dingen in de werkelijkheid en bij het horen van taal. Ook al bestuderen wij het gedrag van mensen en alle taaluitingen die er zijn als we de klank en het object waarnemen, we kunnen nooit bewijzen wat er zich in iemands hoofd afspeelt. Dit is wat hij noemde de '***inscrutability of reference***', de ***ononderzoekbaarheid van verwijzingen***. Dit geldt niet alleen voor mensen die verschillende talen spreken en of culturen bezigen, maar ook voor ieder individu binnen een taalgemeenschap. Deze visie, die teruggaat op de filosofische traditie van de ***sofisten*** en ***nominalisten***, gaat er vanuit dat de werkelijkheid alleen in ons hoofd bestaat. Daartegenover staat de visie van de ***realisten*** die vinden dat de werkelijkheid als een extern object wel bestudeerbaar en kenbaar is. De filosofische discussie hierover is nooit tot een conclusie gekomen, maar dat is misschien ook niet de bedoeling van filosofie.

Voor mijn doeleinden is het echter niet nodig om hier een fundamenteel standpunt over in te nemen. Kijken we naar de drie hoeken of perspectieven van de driehoek, dan zien we dat er vanuit verschillende richtingen veel werk verzet is en nog steeds wordt, met als gevolg dat we veel meer kunnen zeggen over de verankering van de betekenis van woorden. Met andere woorden: de speelruimte voor het probleem is veel kleiner geworden.

2.1 Cognitieve-neurologische modellen van de hersenen

Cognitieve modellen over de werking van onze hersenen in relatie tot perceptie vertellen ons meer en meer over wat er zich in ons hoofd afspeelt, al dan niet met gebruikmaking van steeds betere technieken om hersenactiviteiten te meten. Hoe individueel betekenis ook mag zijn, het kan zich niet onttrekken aan de fysische beperkingen van ons geheugen en onze perceptie en de natuurlijke mechanismen om bewust en zinnig met de waargenomen werkelijkheid om te gaan. De principes rond de vorming van concepten en categorieën kunnen worden beschreven en empirisch worden onderzocht. Nominalisten kunnen zich steeds minder verschuilen achter een blinde vlek omdat de werking van het brein meer en meer als externe werkelijkheid wordt bestudeerd, onafhankelijk van talen en culturen. De inzichten vanuit de neuro-cognitieve hoek beperken dus de speelruimte van betekenis. Een klassiek voorbeeld is de perceptie van kleuren in de vorm van prototypen en graduele verschillen, terwijl de werkelijkheid, het kleuren spectrum, een continuüm is. Als universeel perceptueel gegeven staat het ons toe om te onderzoeken hoe namen voor kleuren zich verhouden tot die perceptuele categorieën en de perceptuele categorieën op hun beurt tot de werkelijkheid. Er zijn talen die bijvoorbeeld maar twee kleurnamen gebruiken (Berlin and Kay 1969, Rosch 1972) en er zijn domeintalen van specialisten die zeer veel namen voor kleuren kennen, vergelijk de honderden kleurentermen in de Getty Art & Architecture

Thesaurus. Ongeacht die verschillen in de lexicalisatie in talen zullen die woorden toch moeten worden ***verankerd*** aan dezelfde perceptuele patronen. Het is de wijze van verankering waarin ik met name in geïnteresseerd ben. Hoe zijn de kleurtermen gerelateerd aan de perceptuele categorieën en wat betekent dat voor het gebruik van die woorden in taal, en uiteindelijk wat zijn de gevolgen voor het denken en communiceren als talen sterk verschillen in lexicalisatie en verankering? Recente studies naar universalia in kleurherkenning en de benoeming van kleuren in 55 niet-industriële talen laten zien dat er zowel universele beperkingen zijn maar ook dat de lexicalisatie in talen van invloed is op de herkenning van niet-focale kleuren (Regier, Kay and Cook 2005). Dit toont precies aan waarom verankering van taal alleen bestudeerd kan worden vanuit beide kanten: cognitie en linguïstiek. Dik (1989) stelt dat als de werkelijkheid helder en eenduidig is, talen dan het cognitieve systeem volgen, maar als de werkelijkheid als vaag wordt ervaren er arbitraire verschillen in de vorm van talen optreden, binnen een taal en tussen talen.

2.2 Formele kennisrepresentatie en ontologieën

Tegenover de cognitieve benadering staat een geheel andere methode vanuit de formele kennisrepresentatie. In de traditie van Aristoteles wordt de werkelijkheid, of de conceptuele representatie daarvan, beschreven door middel van logica. Met formele redeneringen kan de wereld worden ingedeeld in soorten dingen en kunnen de verschillen worden vastgelegd. Een formele definitie van concepten in een kennisrepresentatietaal wordt een ***ontologie*** genoemd. In een dergelijke ontologie wordt bijvoorbeeld een onderscheid gemaakt tussen categorieën als *Physical* en *Abstract* of tussen *Objecten* en *Processen*, waarbij iedere categorie wordt gedefinieerd en van elkaar onderscheiden door middel van een reeks axioma's. De volgende categorieën zijn overgenomen van Sowa (2000):

Abstract (A).

Pure information as distinguished from any particular encoding of the information in a physical medium.:

- No abstraction has a location in space: ~($\exists x$:Abstract)($\exists y$:Place)loc(x,y).
- No abstraction occurs at a point in time: ~($\exists x$:Abstract)($\exists t$:Time)pTim(x,t).

Physical (P).

An entity that has a location in space-time:

- Anything physical is located in some place: ($\forall x$:Physical)($\exists y$:Place)loc(x,y).
- Anything physical occurs at some point in time: ($\forall x$:Physical)($\exists t$:Time)pTim(x,t).

De informatie in deze tekst of de inhoud is bijvoorbeeld *Abstract*. Het bestaat ongeacht de huidige vorm waarin die is gegoten in het hier en nu. Dezelfde inhoud zou al eerder door iemand kunnen zijn bedacht, of misschien dacht u er net zelf aan op dit moment. De tekst zelf, geschreven of gesproken, bestaat in tijd en plaats en kan ook in andere vorm worden verpakt, bijvoorbeeld een schema of diagram. De verschijningsvorm of representatie van de inhoud is dus *Physical*. Dergelijke fundamentele implicaties worden door ons voortdurend gebruikt in ons denken en handelen, bewust en soms onbewust. Het is een logische implicatie die is verankerd in ons bewustzijn en handelen.

Ongeacht de uitkomst van filosofische discussies, is het mogelijk om deze redeneringen te modelleren aan de hand van logische representaties waarin deze categorieën een rol spelen.

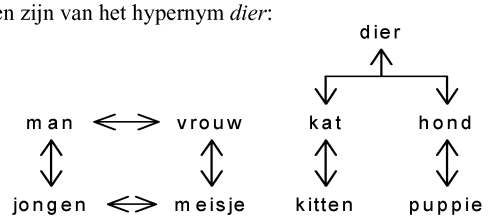
Wat dat betreft wordt er hard gewerkt aan verankering van betekenis in een formeel logisch model. Computerprogramma's van de toekomst zullen aan de hand van deze modellen kunnen redeneren en beslissingen kunnen nemen.

Nu moeten we ons ook niet te veel voorstellen bij een ontologische definitie van werkelijkheden. Volgens een standaard ontologie zoals SUMO (Suggested Upper-level Merged Ontology, Niles & Pease 2001) is een *mobiele telefoon* een subtype van de klasse *Device*, waarbij *Device* gedefinieerd wordt als: an Artifact whose purpose is to serve as an instrument in a specific subclass of Process. Dit zegt niet meer en niet minder dan dat een *mobiel* een soort kunstmatig apparaat is met een bepaalde functie. Het deelt die kennis met duizenden andere apparaten en is er niet verder van onderscheiden omdat er geen verdere informatie wordt gegeven. Het is dan ook niet verwonderlijk dat Quine ontologieën vergelijkt met *woestijnlandschappen.* Ontologieën bevatten alleen de minimale formeel-representeerbare informatie: *ze hebben kraak nog smaak*. De implicaties die in een ontologie zijn vastgelegd zijn abstract. Vanuit dat oogpunt gezien is een ontologie bij lange na niet voldoende om de rijkdom van een taal te beschrijven.

2.3 Linguïstische beschrijvingen van het lexicon en taalgebruik

Wat vaak ontbreekt aan ontologieën is de precieze relatie met natuurlijke taal. Een beschrijving van natuurlijke taal kan zich niet beperken tot alleen die woorden en concepten die formeel definieerbaar zijn. Van woordenboeken wordt verwacht dat de meeste (en in ieder geval al de gangbare) woorden van een taal worden verklaard of geduid. Dat betekent in de praktijk dat alle aspecten van menselijke communicatie moeten worden behandeld en niet alleen de gevallen die het model bevestigen of door het model kunnen worden bevat.

De doelstelling van een woordenboek is de uitleg van woorden aan mensen die die taal al beheersen en een intrinsiek idee van betekenis hebben. Inmiddels is de focus verschoven naar structuren die gebruikt worden door computersystemen, namelijk ***semantische netwerken***. In een semantisch netwerk wordt de betekenis van een woord primair bepaald door de relaties met andere woorden. Die relaties worden niet zozeer vastgelegd in een formele ontologie, maar door woorden met elkaar te vergelijken, al dan niet gebaseerd op het gebruik in teksten. Zonder specifiek te hoeven zeggen wat de relatie nu precies is, kunnen we wel zeggen dat de relatie tussen bijvoorbeeld de woorden *man* en *vrouw* vergelijkbaar is met de relatie tussen *jongen* en *meisje*, even goed kunnen we zeggen dat de relatie *jongen-man* vergelijkbaar is met *meisje-vrouw*, en in zekere zin ook met *kat-kitten* en *puppie-hond* (analoog aan Cruse 1986). Even zo kunnen we stellen dat zowel *kat* en *hond* een soort *dier* zijn net als vele andere dierennamen. Deze laatste relatie wordt een hyponymierelatie genoemd, waarin *kat* en *hond* de hyponiemen zijn van het hypernym *dier*:



Figuur 2: Parallelle impliciete relaties tussen woorden

Er zijn vele soorten semantische relaties tussen woorden die wij in taal aanvoelen. Op deze manier kunnen alle woorden uit een taal ten opzichte van elkaar in kaart worden gebracht. Samen vormen die relaties een zogenaamd ***Wordnet*** (Fellbaum 1998). In een Wordnet representeren woorden met dezelfde betekenis (synoniemen) samen een concept, een zogenaamde *synset*. Zo horen de woorden *man*, *vent*, *kerel* tot dezelfde synset. Synsets worden dan door middel van semantische relaties aan elkaar gerelateerd. We kunnen dit zien als een verankering van woordbetekenis aan elkaar, aan andere woorden. De semantische speelruimte van een woord wordt beperkt door de positie die een woord inneemt in het gehele semantische netwerk van een taal.

Een Wordnet lijkt op een ontologie. De relatie *kat-dier* is namelijk ook terug te vinden in een ontologie. Maar er zijn toch belangrijke verschillen:

- **dekking**: wordnets gaan uit van alle woorden en uitdrukkingen die zijn gelexicaliseerd in een taal en ontologieën richten zich op de concepten die volgens logische principes kunnen worden onderscheiden in de werkelijkheid. Dit levert een zeker spanningsveld op. Niet alle woorden uit talen worden ook gerepresenteerd als categorieën in een ontologie;
- **formaliteit**: wordnets kennen weliswaar verschillende types van semantische relaties en criteria om die te kunnen onderscheiden, maar de relaties zijn minder expliciet geformuleerd in wordnets t.o.v. ontologieën;
- **implicaties**: wordnets kennen geen axioma's die aangeven wat de logische implicaties zijn van die relaties;
- **universaliteit**: wordnets zijn taaleigen en taalintrinsieke systemen terwijl ontologieën universeel pretenderen te zijn;

Er is echter geen fundamenteel onderscheid tussen een wordnet en een ontologie. Zo worden categorieën in een ontologie meestal geschreven met een hoofdletter maar het zijn niettemin (vaak bestaande Engelse) woorden. Het is dan ook niet aannemelijk dat mensen twee representatiesystemen hanteren in hun hoofd: de woorden uit een taal voor communicatie en de categorieën uit een ontologie met een hoofdletter om te denken. Ik zal later terugkomen op de relatie tussen lexicon en ontologie.

Beschrijvingen van natuurlijke taal in de vorm van een grammatica en een lexicon zijn meer en meer gebaseerd op statistieken en validatie van ontzagwekkende verzamelingen teksten of tekstcorpora. Een volledig nieuwe benadering om betekenis te definiëren is gebruik te maken van kwantitatieve informatie. In dat geval wordt er vooral gekeken naar welke woorden er in de buurt staan van andere woorden in zinnen. Zo zouden we uit teksten automatisch kunnen afleiden dat *kat* voorkomt met *miauwen*, *kopjes geven* en *spinnen*. Dit is een hele andere ***verankering*** dan *kat* te associëren met *hond* zoals dat in een semantisch netwerk gebeurt.

Het voert te ver om in deze context hier verder op in te gaan. Het grote verschil tussen de statistische benadering en semantische netwerken en ontologieën is dat de associatie in het eerste geval impliciet blijft. Je weet niet precies wat voor een relatie er is en je kunt het dus ook niet gebruiken om te redeneren over de informatie die je vindt.

2.4 Kunstmatige betekenis voor een computersysteem

Van oudsher hebben mensen nagedacht over de mechanisatie van de mens of het menselijke bewustzijn. De filosofische vraag of denken gemechaniseerd kan worden is opnieuw geactualiseerd binnen de Kunstmatige Intelligentie. Kunnen computers werkelijk intelligent zijn op een menselijke manier? Die vraag is meestal gekoppeld aan het begrijpen en produceren van natuurlijke taal. Turing (1950) bedacht een test aan de hand waarvan men de intelligentie van een computer zou kunnen meten. De zogenaamde Turingtest: iemand wordt in staat gesteld om zonder visueel contact in natuurlijke taal te communiceren met een andere persoon en met een computer. Indien de ondervrager niet kan uitmaken wie de computer is en wie een mens, dan is de computer geslaagd voor de Turingtest. Turing dacht dat het 50 jaar zou duren voor dat computers redelijk zouden scoren op deze test. We zijn nu 50 jaar verder en helaas is het nog lang niet zover.

Voor Alan Turing maakte het niet uit ***hoe*** de computer er in slaagt om de ondervrager 'voor de gek te houden'. Uit deze test volgt niet dat de computer de mens modelleert. De computer hoeft zich niet daadwerkelijk als een mens te gedragen. De Turingtest nodigt uit tot bedrog, het is een zuiver behavioristische test.

Het ultieme bedrog is het programma Eliza van Jospeh Weizenbaum (http://www-ai.ijs.si/eliza/eliza.html). Door het parafraseren van wat iemand intypt weet dit programma de illusie op te wekken dat je praat met een heel geduldige en vriendelijke therapeut. Eliza is geen serieuze poging, het is bedoeld als een parodie. Niettemin werkt het erg suggestief. Er zijn verhalen dat mensen hun ziel en zaligheid hebben verteld aan een computerprogramma dat er niets van begrijpt.

Begin jaren 80 van de vorige eeuw, heeft John Searl de discussie heropend vanuit een nieuw perspectief. Het Chinese Room Argument is een gedachtenexperiment dat door hem is bedacht om aan te tonen dat computers nòòit in staat zullen zijn om natuurlijke taal echt te begrijpen. Het argument gaat uit van de situatie dat iemand in een kamer die geen Chinees spreekt toch in staat is om een adequaat Chinees antwoord te geven door volgens een stel regels de juiste Chinese karakters uit dozen te halen, als een respons op een aantal Chinese karakters die als invoer worden doorgegeven. De regels zijn opgesteld in het Engels (lees een computertaal). Volgens Searl begrijpt de persoon wel de Engelse instructies maar geen Chinees.

In zekere zin hebben we het hier over Eliza, een programma waarvoor de tekens abracadabra zijn en dat toch de verwachte respons genereert. Volgens Searl is er echter geen sprake van 'begrip', sterker nog, volgens Searl kan er fundamenteel nooit sprake zijn van 'begrip'. Het autisme van het systeem, de machine, is zo absoluut dat 'begrip' is uitgesloten. Wat de computer ontbreekt is ***grounding*** in perceptie en bewustzijn. Searl's argument kan worden gezien als een praktisch voorbeeld van Quine's oorspronkelijk argument. In het geval van een Chinese Room kunnen we eigenlijk met zekerheid zeggen dat de conceptuele werkelijkheid van (zeker de huidige generatie) computers anders is dan die van mensen.

Het Chinese Room Argument heeft geleid tot een levendig debat of echt begrip door computers principieel onmogelijk is. Wat volgens mij een veel interessantere vraag is op dit moment is of computers nuttig gedrag kunnen vertonen, waarbij het gaat om de manipulatie van natuurlijke taal in een communicatieve setting. Dat gedrag hoeft helemaal niet intelligent

te zijn, als het ons maar verder helpt. Een programma zoals Eliza is weliswaar bedoeld als parodie, maar wordt door veel gebruikers gezien als een effectieve dialoogpartner waar je je hart kunt luchten terwijl er niet meer gebeurt dan het schuiven met tekstsymbolen zonder enige vorm van begrip of intelligentie.

De enorme groei en dynamiek aan informatie en de snel kleiner wordende wereld, maken hulpmiddelen die ons daarin wegwijs maken steeds belangrijker. Computerprogramma's gaan daar een steeds grotere rol in spelen, ondanks het feit dat ze fundamenteel niets begrijpen. Op dit moment, worden computers en computernetwerken nog meestal gebruikt voor het opslaan en snel doorzoeken van teksten, zoals bijvoorbeeld zoekmachines op het Internet. Deze rol zal echter snel veranderen. Ten eerste zal die informatie steeds beter worden geanalyseerd. Ten tweede gaan computersystemen zelf gebruik maken van die verrijkte informatie en kennis om betere diensten te kunnen leveren.

Teksten kunnen automatisch van labels of metagegevens worden voorzien waar iedereen vervolgens weer nut van heeft. Dit betreft niet alleen wie, wanneer, wat heeft geschreven, maar ook welke *personen* en *producten* erin genoemd worden, wat het *thema* is van de tekst, welke *data* en *plaatsen* erin genoemd worden. Verder kan er automatisch worden doorgelinked naar achtergrond informatie, definities, of andere gerelateerde documenten. De personen die genoemd worden kunnen worden gerelateerd aan *persoonsgegevens*, de *producten* aan *leveranciers* en *afnemers*, *documentatie*, *regelgeving*, enz.. Een tekstdocument staat niet meer op zichzelf maar zal worden verankerd aan allerlei andere informatie die ook weer meer en meer gestructureerd wordt. Op deze manier ontstaat uit het Internet met statische teksten een Informatie of Kennisnet van dynamische objecten die elkaar beïnvloeden. Dit wordt ook wel het Semantic Web genoemd.

De informatie die in een object wordt weergegeven in taal zal dus via de woorden en uitdrukkingen verbonden zijn aan andere objecten die elders staan op het Internet, mogelijkerwijs objecten die de betekenis van die woorden zelf definiëren. Als je een nieuwsartikel leest waarin Balkenende een uitspraak doet, dan is Balkende niet alleen een naam, maar is hij een lid van een politieke partij. Waar die politieke partij voor staat ligt vast in een *verkiezingsprogramma* dat elders op het Internet staat. Zo zou je kunnen stellen dat het *verkiezingsprogramma* een soort definitie is van die *politieke partij*. De informatie in dat verkiezingsprogramma kan worden vergeleken met de uitspraken in het nieuwsartikel. Dat *verkiezingsprogramma* wordt natuurlijk bijgesteld in de loop van de tijd. Een *politieke partij* is voortdurend in beweging, op drift. In een dynamische representatie van kennis en informatie zijn woorden in teksten echter geen statische en vastliggende dingen, maar wordt de informatie waar ze voor staan voortdurend aangepast. Het is dus mogelijk om altijd een actuele definitie van een woord tot je beschikking te hebben. Dat wat een woord betekent, bijvoorbeeld waar een *politiek partij* voor staat, wordt immers elders gedefinieerd.

Maar woorden en objecten zullen ook veel directer met elkaar in verband worden gebracht. Bijvoorbeeld, de definitie wat een bepaald ***type auto*** is, uit welke onderdelen die bestaat, hoe die gemaakt, onderhouden en gerepareerd moet worden, zal meer en meer worden ***voorgeschreven*** vanuit standaarden die worden gedefinieerd in de industrie en/of in regelgeving en wetten van de overheid. De specificaties van die objecten zullen gebruik maken en de vorm hebben van ontologieën en formele kennisrepresentaties. Daarmee wordt de maatschappij een zeer expliciete formele definitie gegeven van een concept, die wettelijke geldigheid heeft. Door dit concept aan woorden te koppelen, wordt het mogelijk voor computerprogramma's om de concepten en hun relaties in teksten te detecteren, verdere

kennis en informatie te vergaren en te benutten. De greep op kennis en informatie vervat in taal zal gestaag groeien.

Naast het structuren en verrijken van de informatie en kennis, zullen computers ook zelf actiever gaan participeren in communicatieve processen. Dit is deels het gevolg van die verrijking zelf, omdat ze beter 'begrijpen' om wat voor informatie het gaat. Een computerprogramma zou bijvoorbeeld zelf actief op zoek kunnen gaan naar ontbrekende informatie, eventueel communiceren met andere computerprogramma's om samen tot een oplossing te komen, bijvoorbeeld de goedkoopste oplossing.

Tenslotte zou een computer op een veel slimmere manier met menselijke gebruikers kunnen interacteren om ze behulpzaam te zijn bij diensten. Een computer zou bijvoorbeeld kunnen assisteren bij het ***kopen*** van een ***auto***. Je overlegt als het ware met een computerprogramma. Het zou de vragen en mogelijkheden of beperkingen van een gebruiker kunnen vergelijken met wat er beschikbaar is. Het zou complexe vragen kunnen beantwoorden door actief deelproblemen op te lossen omdat het weet dat die nodig zijn voor de oplossing van het probleem. Een lening, verzekering, diesel of benzine, etc.... Het zou zelfs kunnen onderhandelen met een aanbieder van auto's, mogelijkerwijs een ander computerprogramma dat actief is voor een andere partij.

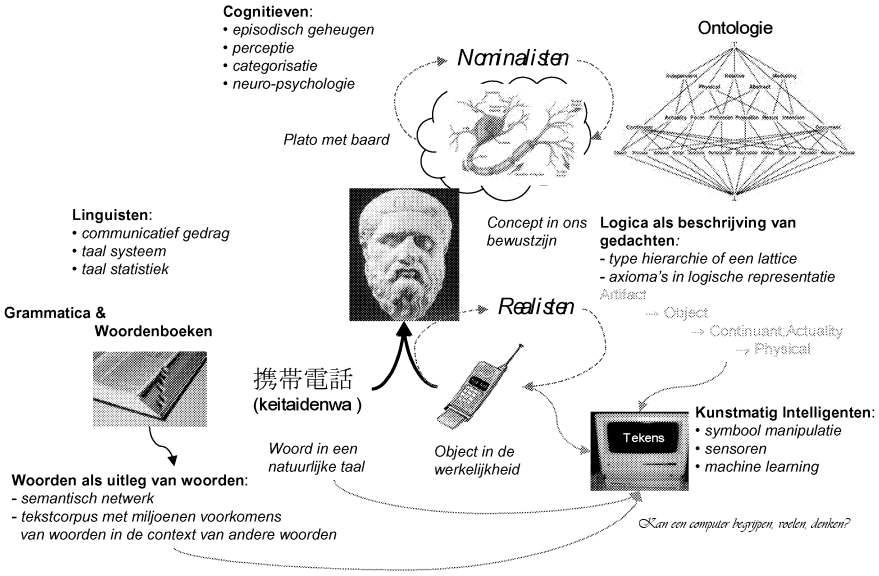
Er mag dan een (haast) onoverbrugbare kloof zijn tussen een eenvoudig programma als Eliza en ware intelligentie, dat wil niet zeggen dat het geen zin heeft om stukjes intelligentie in te bouwen en zo het zinnig hanteren van informatie te verbeteren. Computersystemen zullen in de toekomst een grote rol spelen in de verankering van betekenis. Sterker nog, de kans bestaat dat wij ons begrip van betekenis zullen aanpassen aan het begrip dat voor die computerprogramma's (eventueel wettelijk) is vastgelegd, simpelweg omdat dat het leven makkelijker, aangenamer of productiever maakt. Standaardisatieprojecten nemen een hoge vlucht. De formele definitie van betekenis zal ervoor zorgen dat tekstuele informatie kan worden gezien als een instructie voor computerprogramma's om bepaalde acties te ondernemen. In plaats van dat programmeertalen zich ontwikkelen tot natuurlijke talen, zou het wel eens zo kunnen zijn dat natuurlijke talen zich gaan aanpassen aan de taal die voor een computerprogramma begrijpelijk is en zich zo gaan ontwikkelen dat ze kunnen worden gebruikt als instructietaal voor een computer: d.w.z. tot computertaal. Woorden krijgen voor computers betekenis en daarmee krijgt natuurlijke taal betekenis.

Dergelijke computersystemen combineren linguïstische kennis met ontologische kennis en kennisrepresentaties. Dit zal ze in staat stellen om vrije tekstuele informatie automatisch om te kunnen zetten in kennisobjecten en daarover op een zinnige manier te kunnen redeneren. Het zal ze ook in staat stellen om met menselijke gebruikers in dialoog te gaan over hun informatiebehoefte en ze vervolgens meer op maat te kunnen dienen. Op deze manier wordt betekenis van woorden dus verankerd in functionaliteit die nuttig is. Grounding van betekenis in perceptie met behulp van sensoren en bewustzijn is weer een stap verder. We hoeven dus geen uitspraak te doen of computers daadwerkelijk intelligent kunnen zijn. Het is heel goed mogelijk dat computers allang in staat zijn om kennis en informatie op een efficiëntere manier te hanteren dan mensen voordat ze zelf ook een 'menselijk karakter' zullen krijgen.

Samenvattend kunnen we stellen dat de speelruimte voor woordbetekenis en het mogelijke gebruik van woorden in teksten zowel aan de conceptuele als aan de linguïstische kant flink beperkt wordt door de ontwikkelingen die plaatsvinden. Woorden zijn niet op drift maar kunnen tot op zekere hoogte worden gekoppeld aan formeel gedefinieerde concepten, aan

universele cognitieve modellen en aan elkaar, hetzij in een semantisch netwerk of in combinatorisch gedrag in teksten. Daarnaast zien we dat computersystemen steeds meer gebruik gaan maken van de beschikbare informatie en kennis die uit teksten kan worden gedestilleerd. Vervolgens is het niet meer zo'n kleine stap om ook de vragen en uitleg van gebruikers van computersystemen te begrijpen en die om te zetten in instructies die moeten leiden tot nuttig gedrag.

Zo gezien, staat Plato niet meer alleen met zijn gedachten maar kan er veel worden ingevuld. In Figuur-3 staat nu een overzicht van de eerdere driehoek, waarin is aangegeven hoe vanuit iedere hoek een verankering kan plaatsvinden van de verschillende relaties tussen concept , woord, en werkelijkheid.



Figuur 3: Verschillende paradigma die invulling geven aan de relatie concept, woord en object.

In het vervolg wil ik ingaan op wat mijn bijdrage is aan het verankeren van woordbetekenis in dit plaatje. Dit betreft op dit moment drie activiteiten:

1. De ***Global Wordnet Grid***: een wereldwijd project van de Global Wordnet Association dat ik gestart heb om alle talen in de wereld te verankeren aan een universele concept index;
2. Het Stevin project ***Cornetto*** waarin de relatie tussen een Nederlands lexicon en een ontologie wordt bestudeerd. Het lexicon bevat zowel een semantisch netwerk (*kat-hond-dier*) als combinatorische informatie (*kat – kopjes geven*) van woorden, maar wordt ook gekoppeld aan een formele ontologie;
3. ***Informatiedialogen*** en ***kennisontginning*** bij Irion Technologies, waarbij aan de ene kant informatie wordt gekoppeld aan concepten en informatiebehoeften en aan de

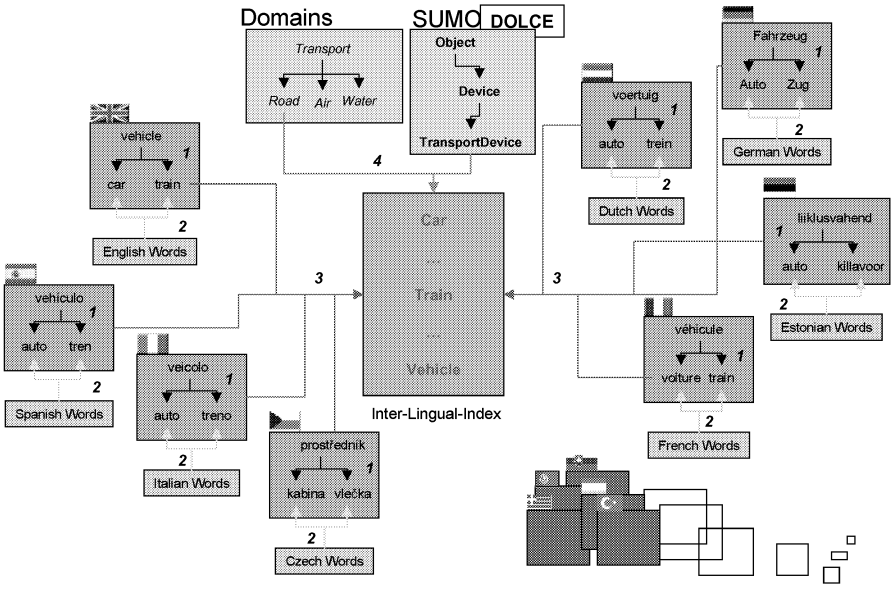
andere kant een linguïstisch dialoogsysteem gebruikers vraagt naar wat ze willen en ze vervolgens informatie op maat aanbiedt;

In de volgende 3 paragrafen zal ik hier verder op in gaan om aan te geven hoe de verankering van woordbetekenis hierbinnen een rol speelt.

3 Global Wordnet Grid

Een semantisch netwerk of wordnet is een belangrijke datastructuur die zowel een taalcultuur karakteriseert, als weergeeft op een conceptueel-cognitief niveau welke concepten en relaties wij kennen. De woorden van een taal zijn het resultaat van eeuwen aan conceptvorming, cultuur en historie. Ze vormen een kaart van ons denken, voelen en doen binnen een gemeenschap.

Het Engelse WordNet werd als eerste ontwikkeld door George Miller en zijn groep in Princeton University als een implementatie van het mentale lexicon (Fellbaum 1998). In 1996 is dit model verder uitgebreid in het EuroWordNet project (Vossen 1998). In dit EU-project (LE-2 4003 & LE-4 8328) zijn wordnets gemaakt voor 7 andere Europese talen maar is bovendien een model gemaakt waarbij die wordnets met elkaar verbonden zijn door middel van een zogenaamde Inter-Lingual-Index (ILI). Deze ILI is een voorraad aan concepten die uitsluitend dienst doen om equivalente concepten uit talen aan elkaar te verbinden. Het kan dus gezien worden als een universele betekenisindex. Binnen de architectuur van EuroWordNet is ieder wordnet een unieke verzameling concepten en relaties in iedere taal, maar tegelijkertijd is het mogelijk om vanuit ieder concept in een taal naar equivalente concepten in iedere andere taal te gaan via de ILI, de universele index. Op die manier kunnen de lexicalisaties van talen met elkaar worden vergeleken en zijn meertalige oplossingen te maken zoals zoeken in meerdere talen tegelijk:



Figuur 4: Architectuur van de EuroWordnet database

In figuur-4 worden vier soorten relaties onderscheiden:

1. Taal-interne semantische relaties tussen de concepten die in een taal gelexicaliseerd zijn (zwarte pijlen (1) tussen de concepten in ieder wordnet);
2. Relaties tussen woorden en concepten (pijlen gemarkeerd met 2). Woorden kunnen naar meerdere concepten verwijzen en dus meerdere betekenissen hebben (polysemie). Andersom kunnen meerdere woorden dezelfde betekenis hebben, oftewel synoniem zijn;
3. Relaties tussen de concepten van een taal en de ILI (de pijlen gemarkeerd met 3). Dit worden ook wel equivalentie relaties genoemd;
4. Relaties tussen taalonafhankelijke ontologieën en thesauri met domeinaanduidingen en de ILI (pijlen gemarkeerd met 4);

Binnen het EuroWordnet model is het mogelijk om een principiëler standpunt in te nemen welke concepten en relaties in een wordnet moeten worden uitgedrukt. Welke concepten moeten worden opgenomen in een Wordnet is een zaak voor iedere taal. Hetzelfde geldt voor de semantische relaties tussen die woorden. Wel worden die taal-interne systemen aan elkaar verankerd door middel van de ILI. De semantische implicaties die in het ene wordnet worden uitgedrukt zullen dus toch moeten passen in of harmoniseren met de relaties die worden uitgedrukt in alle andere wordnets. Als *trein* een *voertuig* is in het Nederlandse wordnet dan verwachten we dat het equivalent in andere talen dezelfde relatie heeft met het equivalent van *voertuig* in die talen.

Na het EuroWordnet project zijn er wereldwijd meer en meer initiatieven geweest om wordnets te ontwikkelen voor talen in de wereld. Op dit moment zijn er meer dan 40 wordnets geregistreerd op de Global Wordnet website (http://www.globalwordnet.org):

> *Afrikaans; Arabic; Bantu-talen; Basque; Bulgarian; Catalan; Chinese; Czech; Danish; Dutch; English; Estonian; French; German; Greek; Hebrew; Hindi; Hungarian; Icelandic; Italian; Kannada; Korean; Latijn; Latvian; Marathi; Moldavian; Norwegian; Oriya; Persian; Portuguese; Brazillian Portuguese; Romanian; Russian ; Sanskrit; Serbian; Slovenian; Spanish; Swedish; Tamil; Thai; Turkish*

Om het raamwerk van EuroWordnet te kunnen continueren na het project heb ik samen met Christiane Fellbaum van Princeton University de Global Wordnet Association (GWA) opgericht. Deze stichting bestaat nu 6 jaar en heeft 3 succesvolle internationale congressen georganiseerd, de vierde is op komst. De doelstelling van GWA is de ontwikkeling van wordnets in alle talen van de wereld te steunen en te stimuleren, maar ook om te zorgen voor compatibiliteit van die wordnets, met als ultieme doel een wereldwijd raamwerk van wordnets: de zogenaamde Global Wordnet Grid.

De bedoeling van de Global Wordnet Grid is niet alleen het koppelen van alle talen in de wereld aan elkaar maar ook met name het sturen van de discussie over lexicale universalia. Welke concepten worden door veel of alle talen gedeeld, zijn er systematische verschillen in de lexicalisatie van concepten tussen talen? Hoe wordt een concept gedefinieerd, wat is de verzameling concepten wereldwijd, wat zijn de woorden en uitdrukkingen in een taal die moeten worden opgenomen? Hoe gaan we om met verschillen in lexicalisaties en verschillen in relaties tussen die concepten in multilinguale contexten? We hebben gezien dat kleuraanduidingen enorm kunnen verschillen tussen talen, terwijl de fysische werkelijkheid (het spectrum) en onze perceptie daarvan universeel bepaald is. Wat moet dat dan niet

betekenen voor andere woorden en begrippen die veel grilliger zijn en meer verankerd zijn in een culturele werkelijkheid?

In het bovenstaande model worden alle talen dus aan elkaar verankerd door middel van de universele conceptindex, de ILI. De ILI is nu gevuld met de concepten uit het Engelse wordnet, zoals in figuur-4 aangegeven met *car*, *train*, en *vehicle*. Dit is voornamelijk gedaan uit praktische en pragmatische redenen. Toentertijd was er geen andere bron vrij beschikbaar die zoveel concepten bevatte, bovendien beheersen de meeste mensen die wordnets ontwikkelen tot op zekere hoogte het Engels. Het gevolg is echter wel dat het koppelen van betekenissen wordt bemoeilijkt door een culturele en linguïstische bias. Specifieke talige eigenschappen van het Engels zitten in de weg. Bovendien ontbreken er concepten die in andere talen voor de hand liggen maar in het Engels niet gelexicaliseerd zijn. De vraag is nu of het Engelse Wordnet wel een goede kandidaat is voor de verankering van de verschillende talen. Hieronder volgen een paar voorbeelden ter illustratie van de problemen die optreden bij het vinden van een equivalent Engelse concept in de ILI.

Culturele gaps
Sommige concepten komen niet voor in het Engelse WordNet omdat ze niet of nauwelijks in de Anglo-Saxische cultuur voorkomen. Dit soort woorden vormen zogenaamde *cultural gaps* in het Engelse WordNet. Deze concepten kunnen uitgelegd worden maar niet vertaald. Een voorbeeld is *klunen*: “to walk on skates over land from one frozen water to the next, during skating”. Een ander voorbeeld is *citroenjenever* dat niet gemaakt wordt in de Anglo-Saxische cultuur en dus niet gevonden in WordNet.

Pragmatische gaps
Veel frequenter zijn zogenaamde pragmatische gaps (Vossen, Peters, and Gonzalo 1999). Dat zijn woorden voor concepten die wel bekend of herkend worden in een andere cultuur/taal maar niet als gelexicaliseerd worden beschouwd. In talen zoals het Nederlands worden bijvoorbeeld veel samenstellingen als woorden beschouwd waarbij dat in de Engelse taalcultuur niet het geval is, bijvoorbeeld *mensenrechtenactivist*. Vaak is de vorming van deze woorden productief en regelmatig, zoals bij Nederlandse werkwoorden waarbij het resultaat of de voortgang en de manier waarop worden gecombineerd:

- *doodslaan* (beat to death), *doodtrappen* (kick to death), *doodknijpen* (squeeze to death), etc.
- *fijnslaan* (beat to become crushed), *fijntrappen* (kick/stamp to become crushed), *fijnknijpen* (squeeze to become crushed), etc.

De betekenis is grotendeels systematisch. Al deze woorden kunnen in het Engels worden geparafraseerd door twee werkwoorden of een werkwoord dat de manier aangeeft en een constructie die het resultaat of de voortgang weergeeft. Hier speelt de vraag een rol wat nu precies een woord is. Representeert iedere mogelijke samenstelling in iedere taal een apart woord en een apart concept dat als zodanig ook in de ILI moet worden opgenomen?

Linguïstische gaps
Als woorden grammaticale aspecten incorporeren is de relatie met de andere taal moeilijker te maken. In Slavische talen zoals het Tsjechisch worden bijvoorbeeld aspectuele verschillen door middel van verschillende werkwoorden uitgedrukt (voorbeelden uit Õim, Vider, Paldre, Orav, and Pala 1999).

Imperfectief	*Bijna equivalent in Engels*	*Perfectief*	*Parafrase in Engels*
brat; vzít	*take*	*brát si, brát se*	*take for one's benefit : take and have it,*
d• lat	*do*	*ud• lat*	*finish doing*
dávat	*keep giving*	*dát*	*give and finish giving*

Hier is het Engelse werkwoord steeds algemener en moet uit de context blijken wat de fase is van een proces (is het nog gaande, is het compleet) voordat je het juiste Tsjechische werkwoord kunt kiezen. In het Nederlands zien we een vergelijkbaar verschijnsel als werkwoorden worden gecombineerd met voorzetsels of andere werkwoorden die een aspect impliceren:

- *doorslaan*[i] (continue to beat), *doortrappen* (continue to kick), *doorknijpen* (continue to squeeze), etc.

Deze specifieke aspectuele lezingen kunnen niet precies aan de ILI worden gekoppeld omdat het Engelse werkwoord te algemeen is.

Een andere klasse wordt gevormd door verschillen in biologisch geslacht vaak uitgedrukt in morfologische structuur, vergelijk:

- theoloog/theologe (theologist)
- leraar/lerares (teacher)
- dokter/arts (doctor)

De Engelse woorden *theologist*, *teacher*, en *doctor* zijn allemaal neutraal wat betreft man of vrouw. Een *theoloog* in het Nederlands is ook neutraal maar een *theologe* niet. Een *leraar* daarentegen is altijd een man en een *lerares* een vrouw. Er is dus geen neutrale betekenis. Een *arts/dokter* is altijd neutraal en er is geen vrouwelijke vorm: *artsin* is geen legitiem woord. Om de betekenisverschillen van deze subtiele variaties in de ILI weer te geven voor alle woorden die voor geslacht markeren, moeten we dus naast de neutrale Engelse betekenissen ook mannelijke en vrouwelijke varianten opnemen.

Een vergelijkbaar verschil wordt gevormd door Nederlandse verkleinvormen zoals *kopje* en *glaasje* die in het Engels nooit gelexicaliseerd zijn maar in het Nederlands soms pragmatisch geladen zijn (meer beleefd, niet gulzig) en dan een legitiem woord vormen.

In zoverre dergelijke verschillen tussen talen (aspect, geslacht, verkleinvormen, e.d.) systematisch zijn zouden ze de ILI enorm opblazen indien we ze allemaal opnemen.

Inclusie verschillen
In veel andere gevallen wordt het verschil niet bepaald door een morpho-syntactisch verschijnsel, maar wordt er een bepaalde semantische implicatie in het Engels niet gemaakt. Het Engelse werkwoord *cut* bijvoorbeeld laat in het midden hoe de snede tot stand is gekomen, terwijl wij in het Nederlands gedwongen zijn een onderscheid te maken tussen *knippen, snijden* en *hakken.* De gangbare Nederlandse woorden hebben dus een sterkere inclusie dan het Engels en kunnen dus niet direct aan de ILI worden gekoppeld. We kunnen hooguit *snijden* en *hakken* als specifiekere concepten aan *cut* hangen maar de relatie met het

instrument en de precieze beweging wordt daarmee niet duidelijk gemaakt. Er is dan alleen een verschil in specifiekheid of inclusie tussen andere talen en het Engels, zonder dat daar een duidelijke motivatie voor is.

Dit is slechts een kleine greep uit het scala aan mismatches tussen concepten uit andere talen met het Engels, waarbij ik grotendeels naar een verwante taal als het Nederlands heb gekeken en niet naar exotische talen. Een van de kernvragen die zich opdringt is wat we precies onder een concept verstaan in deze context. Welke concepten zouden we moeten opnemen in de universele index? Simpelweg stellen dat ieder woord een apart concept is is geen oplossing. Het is niet eens duidelijk wat een woord is. Sommige talen gebruiken een uitdrukking waar andere talen een woord zonder spaties gebruiken. Verder zijn er allerlei varianten van woorden, bijvoorbeeld meervoud en enkelvoud, verkleinwoorden, afkortingen, etc. die we niet als aparte concepten zouden willen zien, evenals synoniemen. Kortom, het is nodig dat er een woordonafhankelijke definitie komt wat een concept is en daaraan gekoppeld welke woorden of woordvormen tot hetzelfde concept horen en dus varianten of synoniemen zijn.

Voor de ontwikkeling van een Global Wordnet Grid is daarom voorgesteld om een taal-onafhankelijke ontologie als de ILI te nemen. Op dit moment wordt gedacht aan SUMO (Niles and Pease 2001), met name omdat het vrij beschikbaar is en een redelijke dekkingsgraad heeft. SUMO is het resultaat van het samenvoegen en bewerken van een aantal beschikbare ontologieën. Bovendien zijn alle synsets in WordNet voorzien van een SUMO concept. Dat betekent dat het in principe mogelijk is om de huidige relaties naar de ILI, automatisch over te zetten naar relaties naar SUMO.

Een ontologie als een ILI heeft een aantal voordelen boven het Engelse WordNet:

- Er is minder of geen bias vanuit de talige structuur;
- Er wordt zuiniger omgegaan met het definiëren van concepten, er is minder wildgroei;
- Er is een meer expliciete definitie van wat een concept is en waarin het zich onderscheid van andere concepten;
- Het kan worden gebruikt om te redeneren ongeacht de specifieke lexicale realisatie in talen en op uniforme manier gebaseerd op logica;
- Het is mogelijk om complexe concepten af te leiden aan de hand van kennisrepresentatie formalismen zoals het Knowledge Interchange Format (KIF). Een KIF expressie is een vereenvoudigde representatie of formule voor het uitdrukken van logische statements. Hiermee kunnen bepaalde taalspecifieke concepten worden gedefinieerd zonder de ontologie te moeten uitbreiden;

Het belangrijkste aspect van een ontologie is dat het een meer fundamentele definitie geeft van wat concepten zijn. Klassen van entiteiten kunnen worden onderscheiden op grond van zogenaamde identiteitscriteria (Guarino en Welty 2002). Hierin spelen de volgende aspecten een rol:

- **rigiditeit**: in wat voor mate zijn eigenschappen van entiteiten waar in alle mogelijke werelden? Een *mens* ben je altijd, een *student* kun je tijdelijk zijn.
- **essentie**: welke eigenschappen zijn essentieel voor een entiteit? Vorm is essentieel voor een *beeld* en niet voor de *klei* waaruit een beeld bestaat.
- **uniciteit**: wat vormt een geheel en welke dingen zijn een onderdeel van een geheel? Een *zee* is een geheel maar *water* niet.

Rigiditeit bepaalt het onderscheid tussen disjuncte typen en zogenaamde rollen. Een *kat* en een *hond* zijn rigide en daarom disjuncte klassen. Iets kan niet zowel een *kat* als een *hond* zijn, en ook niet eerst een *kat* en dan een *hond*. Een *huisdier* is echter antirigide. Een dier kan eerst *wild* zijn en dan een *huisdier* worden en vervolgens weer *verwilderen*. Hetzelfde geldt voor *reiziger* of *passagier*. Zodra je op de bus stapt ben je een *passagier* maar zodra je uitstapt niet meer. Begrippen als *huisdier, student* of *reiziger* zijn rollen en niet soorten dingen.

Guarino en Welty beschrijven een methode, genaamd OntoClean, om deze criteria toe te passen. Het gaat hier te ver om de precieze definitie en methode verder uit te leggen. Op grond van deze noties komen ontologisten tot een indeling van de werkelijkheid die een meer zuivere en compacte verzameling is van alleen de disjuncte typen met daarnaast een definitie van eigenschappen en rollen die op die types van toepassing kunnen zijn. Zij hebben deze methode ook toegepast op het Engelse WordNet. Daarbij laten ze zien dat sommige relaties niet kloppen of niet precies duidelijk maken wat de semantische implicatie is.

Er zijn ook nadelen verbonden aan het gebruiken van een ontologie als een ILI. Het grootste nadeel is de dekking. SUMO (samen met de uitbreiding MILO) kent weliswaar 20,000 concepten maar het Engelse WordNet 115,000. Dat betekent dat de meeste synsets in WordNet naar een ontologische klasse als subtype verwijzen die veel algemener is, zoals we gezien hebben met *mobiele telefoon* dat geclassificeerd is als een type ***Device***.

Bij een systeem waar het Engels de interlingua is zien we dat het soms erg moeilijk is om het juiste concept te kiezen om synsets in een taal te relateren aan de ILI. In EuroWordNet is gebleken dat geregeld andere keuzes worden gemaakt vanuit verschillende talen tussen fijnmazige nuanceringen in het Engels. Nemen we een ontologie zoals SUMO als de interlingua dan krijgen we het omgekeerde probleem. Er zullen duizenden woorden zijn in iedere taal die gekenmerkt zijn als een type ***Device***. Op grond daarvan zullen we nooit kunnen bepalen welke woorden nu de equivalenten zijn tussen talen. We verliezen dus zeggingskracht.

Er moeten een aantal dingen gebeuren wil een ontologie als SUMO geschikt worden als ILI, waarvan de twee belangrijkste zijn:

1. De disjuncte oftewel rigide types (zoals soorten Devices) moeten worden uitgebreid zodat een dekking wordt bereikt die minstens equivalent is aan het Engelse WordNet;
2. De antirigide concepten -- pragmatische woorden, inclusie verschillen, e.d. – moeten door middel van een KIF expressie worden gedefinieerd;

Dit zou kunnen gebeuren door de concepten in SUMO op te blazen met de rigide types uit het Engelse WordNet. Helaas volgt uit WordNet niet welke woorden disjuncte types zijn en welke woorden dat niet zijn.[ii] Kijken we bijvoorbeeld naar alle subtypes (hyponiemen) van *dog* in het Engelse WordNet dan vinden we naast *poodle*, *Newfoundland*, en *dalmatian* ook honden als *lapdog*, *hunting dog*, en *working dog*. De eerste zijn duidelijk rigide: eenmaal geboren als een *poodle* wordt je nooit meer een *Newfoundland*, en de tweede groep zijn honden in een rol die tijdelijk kan zijn of kan veranderen, d.w.z. *antirigide*.

In een ontologie wordt alleen een onderscheid gemaakt tussen de rigide hondenrassen:

Canine => PoodleDog; NewfoundlandDog; DalmatianDog, etc.

Hierbij heb ik de Engelse namen met hoofdletters geschreven om aan te geven dat het hier om disjuncte typen dieren gaat (klassen van objecten in de werkelijkheid) en niet om de woorden in een taal. De Engelse woorden kunnen dan direct gelijkgesteld worden aan deze types. Het Engelse woord *poodle* is dan de Engelse naam voor het disjuncte type *PoodleDog*.

Voor de andere Engelse *honden*-woorden hoeven we dus niet de ontologie uit te breiden maar kan een SUMO-KIF expressie worden gebruikt (Pease 2000) waarin duidelijk wordt gemaakt dat het om een (willekeurige) *hond* gaat in een bepaalde rol. Een dergelijke expressie zou er dan als volgt uitzien:

```
hunting dog (?CAN)
        ⇨  (exists (?CAN ?EV)
                (and
                        (instance ?CAN Canine)
                        (instance ?EV Hunting)
                        (agent ?CAN ?EV)))
```

In deze notatie zijn de woorden voorafgegaan door een vraagteken variabelen voor mogelijke entiteiten in de wereld. In gewone taal staat hier dus dat een entiteit ?CAN die je een *hunting dog* noemt in het Engels, een instantie is van het type Canine en dat er dan ook een instantie bestaat van het proces *Hunting*, zodanig dat ?CAN de agens is in dit proces. In dit geval wordt door middel van een axioma niet een subtype relatie gelegd maar wordt de rol expliciet aangegeven in het logische frame. De ontologie hoeft dus niet te worden uitgebreid met een nieuw type *HuntingDog*.

Het proces van het onderscheiden van rigide typen en het definiëren van rollen zal dus grotendeels met de hand moeten gebeuren door specialisten. Is het onderscheid eenmaal gemaakt in het Engelse WordNet, dan kan SUMO worden 'opgeblazen' tot een volledige lijst van disjuncte typen en kunnen ook alle andere talen hiervan profiteren. Woorden die namen zijn voor rigide klassen kunnen dan direct gekoppeld worden aan de corresponderende types in de ontologie:

MobilePhoneDevice (nieuw rigide type in SUMO)
= mobiele telefoon, gsm, mobiel (NL)
= 携帯電話 (*keitaidenwa, keitai*) (JP)
= mobile phone, cell phone, cell (EN)

Indien een taal een woord heeft voor een rigide type dat niet bestaat in het Engels, zoals voor de Cultural Gaps *klunen* of *citroenjenever*, dan zal dat type moeten worden toegevoegd aan de ontologie met de relevante axioma's. De ontologie zal dus uiteindelijk alle rigide types bevatten die door alle culturen en talen worden bijgedragen. Gelukkig zijn de meeste en de meest-problematische verschillen tussen talen niet *cultural gaps* maar eerder pragmatisch en linguïstisch van aard. Hiervoor hoeft in principe de ILI niet te worden uitgebreid maar kunnen we volstaan met SUMO-KIF expressies.

De Nederlandse samenstellingen *doodslaan* en *fijnslaan* die als pragmatische gaps in het Engels ontbreken, kunnen we bijvoorbeeld volledig definiëren met een KIF expressie die gebruik maakt van andere ontologische concepten *Die*, *Hit* en *Attack*:

doodslaan

⇨ doden, slaan (wordnet relaties)
⇨ (exists (?PROC1 ?PROC2)
(and
(instance ?PROC1 Attack)
(instance ?PROC1 Hit)
(instance ?PROC2 Die)
(causes ?PROC1 ?PROC2)))

Dezelfde expressie kan gebruikt worden voor het Duitse woord *totslagen*. Equivalentie tussen het Duits en Nederlands zal dan indirect blijken uit het feit dat de woorden gekoppeld zijn aan dezelfde formule. Deze oplossing heeft drie belangrijke voordelen:

- We stellen hier expliciet dat deze werkwoorden antirigide en niet-disjuncte typen processen zijn;
- We hoeven de ILI niet uit te breiden met duizenden 'pragmatische' concepten om equivalentie tussen talen uit te drukken;
- We geven in een formele representatie aan wat de relatie is tussen de subprocessen die hier worden genoemd, namelijk dat het ene proces het andere veroorzaakt;

Aan de hand van deze uitdrukkingen kunnen we ook andere gevallen oplossen die hierboven genoemd zijn. Beroepen met een man/vrouw variant kunnen worden gerelateerd aan neutrale beroepsrollen in combinatie met het attribuut mannelijk of vrouwelijk of een instance relatie met het SUMO concept ***Man*** of ***Woman***:

lerares

⇨ vrouw, agens-in-lesgeven (wordnet relaties)
⇨ (exists (?HUM ?PROC)
(and
(instance ?HUM Woman)
(attribute ?HUM SkilledOccupation)
(instance ?PROC EducationalProcess)
(agent ?HUM ?PROC)))

leraar

⇨ man, agens-in-lesgeven (wordnet relatie)
⇨ (exists (?HUM ?PROC)
(and
(instance ?HUM Man)
(attribute ?HUM SkilledOccupation)
(instance ?PROC EducationalProcess)
(agent ?HUM ?PROC)))

Hierbij moet worden opgemerkt dat in SUMO rollen als *Attributen* worden opgeslagen:

SkilledOccupation=>Position=>SocialRole=>RelationalAttribute=>Attribute=>Abstract.

Een rol wordt dus niet gezien als een soort ding maar als een relatie.

In het geval van een regelmatig woord zoals *theoloog* dat neutraal of mannelijk kan zijn, kunnen we volstaan met één KIF-expressie, net als voor *arts/dokter* dat altijd neutraal is.

Ten slotte, zowel het Engels als SUMO definiëren het algemenere proces *cutting*, zonder implicatie van het instrument dat wordt gebruikt. We kunnen dan de specifiekere concepten *snijden* en *knippen* echter wel definiëren aan de hand van een KIF expressie, waarin we aangeven dat de een met een *mes* en de ander met een *schaar* moet plaatsvinden:

```
snijden
        ⇨ (exists (?INSTR ?PROC)
                (and
                        (instance  ?INSTR Knife)
                        (instance ?PROC Cutting)
                        (instrument ?INSTR ?PROC)))
knippen
        ⇨ (exists (?INSTR ?PROC)
                (and
                        (instance  ?INSTR Scissors)
                        (instance ?PROC Cutting)
                        (instrument ?INSTR ?PROC)))
```

Hier kan echter wel de vraag gesteld worden of de ontologie niet te abstract is (onder invloed van het Engels). Men zou goed kunnen beargumenteren dat *snijden* en *knippen* twee heel andere rigide processen zijn. In dat geval zou de ontologie dus eigenlijk moet worden uitgebreid met een *CuttingWithKnife* en *CuttingWithScirrors* proces en is het Engelse werkwoord *cut* daar een abstractie van.

Door bovenstaande strategie te volgen kan de ontologie compact en exact worden gehouden terwijl de expressiviteit voldoende blijft om iedere betekenis uit iedere taal weer te geven. In potentie zou een ontologie dus kunnen dienen als een ILI om de woordbetekenissen van alle talen aan te koppelen. Voor dat het zo ver is moet echter nog veel werk verzet worden. Voor het Nederlands wordt dat op dit moment gedaan binnen het Stevin project Cornetto dat nu wordt uitgevoerd en dat ik nu namens de Vrije Universiteit in het kader van deze leerstoel heb aangevraagd en coördineer.

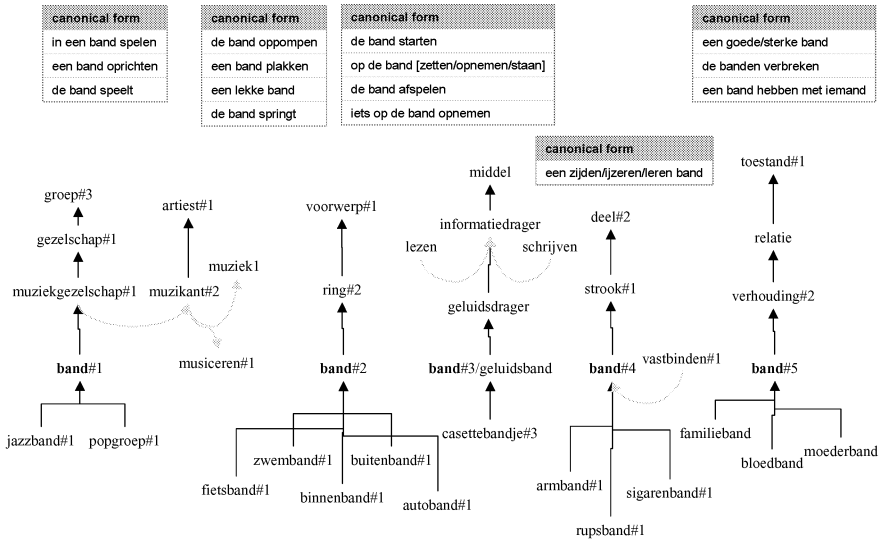
4 Cornetto: lexicon, wordnet en ontologie

Het Cornetto project (STE05039) wordt gefinancierd door het Stevin programma. De partners in het project zijn de Vrije Universiteit van Amsterdam, de Universiteit van Amsterdam, de Katholieke Universiteit Leuven en een taaltechnologiebedrijf Irion Technologies. Het is een 2-jarig project dat als doel heeft een lexicale database voor het Nederlands te bouwen met zowel semantische relaties als combinatorische informatie die nodig is om woorden te combineren.

De methode die in Cornetto wordt gehanteerd is het samenvoegen van twee bestaande databases, namelijk het Nederlandse Wordnet (Vossen 1998) en het Referentie Bestand Nederlands (Maks, Martin en Meerseman 1999) en die verder te verbeteren door de woordbetekenissen te koppelen aan een formele ontologie, min of meer zoals hierboven beschreven.

Het Nederlandse wordnet (DWN) is vergelijkbaar met het Engelse WordNet en bevat verticale en gedeeltelijk ook horizontale semantische relaties. Het Referentie Bestand Nederlands (RBN) is ook semantisch van opzet maar bevat uitsluitend horizontale relaties en met name combinatorische informatie. Vind je in het DWN dat *koffie* een soort *drank* is en wellicht dat *dranken gemaakt worden* dan vind je in het RBN dat je *koffie en thee zet*[iii] terwijl je *limonade maakt*. Hetzelfde geldt voor voorzetsels bij werkwoorden zoals *behandelen*. Vind je in het DWN dat er een semantische relatie is tussen *behandelen* en *ziektes* en *verwondingen*, dan vind je in het RBN dat die relatie in het Nederlands gerealiseerd wordt met verschillende voorzetsels: *behandelen aan zijn verwondingen* maar *behandelen voor een ziekte.* Er is een heel scala aan dergelijke combinatorische informatie dat typisch is voor het Nederlands.

Door het RBN met het DWN te combineren ontstaat een zeer rijk netwerk van conceptuele én combinatorische informatie. Deze informatie is essentieel voor computers om de betekenis van woorden in teksten te herkennen, maar ook om vloeiende teksten te genereren in toepassingen. In figuur-5 hieronder worden bijvoorbeeld een aantal betekenissen van het woord *band* weergegeven in verschillende stukjes wordnet-relaties. De zwarte pijlen naar boven geven subtypenrelaties aan en de lichtere gebogen pijlen andere relaties. Het gaat hierbij om de betekenissen: *muziekband*, band als *stukje materiaal*, een *emotionele/psychische* band en een band als *medium voor opnames*.



Figuur 5: Wordnet fragmenten en combinatorische informatie voor verschillende betekenissen van *band*

Boven ieder fragment wordt in een tabel de combinatorische informatie gegeven uit het RBN die met deze betekenissen correspondeert. In deze woordcombinaties herkennen we meteen de betekenissen uit het Nederlandse wordnet.[iv] Het interessante aan de combinatie van DWN en RBN is bovendien dat de combinatorische informatie uit het RBN kan worden

gegeneraliseerd naar de specifieke soorten *banden* en zelfs voor sommige van de algemenere woorden, vergelijk:

- ijzeren rupsband
- in een popgroep spelen
- een muziekgezelschap oprichten
- iets opnemen op een geluidsdrager
- een lekke binnenband, buitenband
- een sterke moederband
- een sterke verhouding

De combinatie van RBN en DWN representeert twee verschillende manieren van organisatie van linguïstische kennis die elkaar aanvullen en versterken. Het is daarvoor wel noodzakelijk dat de woordbetekenissen van beide bronnen op de juiste manier aan elkaar gekoppeld worden. Deze koppeling wordt verder ondersteund en verrijkt door iedere betekenis aan een ontologie te koppelen.

4.1 De architectuur van de Cornetto database

In Cornetto gaat het niet om woorden maar om woordbetekenissen. We gaan er vanuit dat woorden altijd gebruikt worden in een bepaalde betekenis en dat het geen zin heeft om de vorm van een woord los te zien van die betekenis. Het woord *meer* als bijvoeglijk naamwoord staat los van het zelfstandig naamwoord. Dat neemt niet weg dat er relaties kunnen zijn tussen verschillende woordbetekenissen.

Binnen Cornetto proberen we een expliciete definitie te geven van woordbetekenissen door het leggen van relaties tussen:

- de vormeigenschappen van een woord;
- de wijze waarop een woord combineert met andere woorden;
- de semantische plaats/positie die het woord heeft binnen het Nederlandse wordnet;
- de relatie die het woord heeft met een formele ontologie;

Een woordbetekenis wordt ook wel een Lexical Unit (LU) genoemd (Cruse 1986). Een woordbetekenis of LU wordt dan gedefinieerd als een vorm-betekenisrelatie zodanig dat:

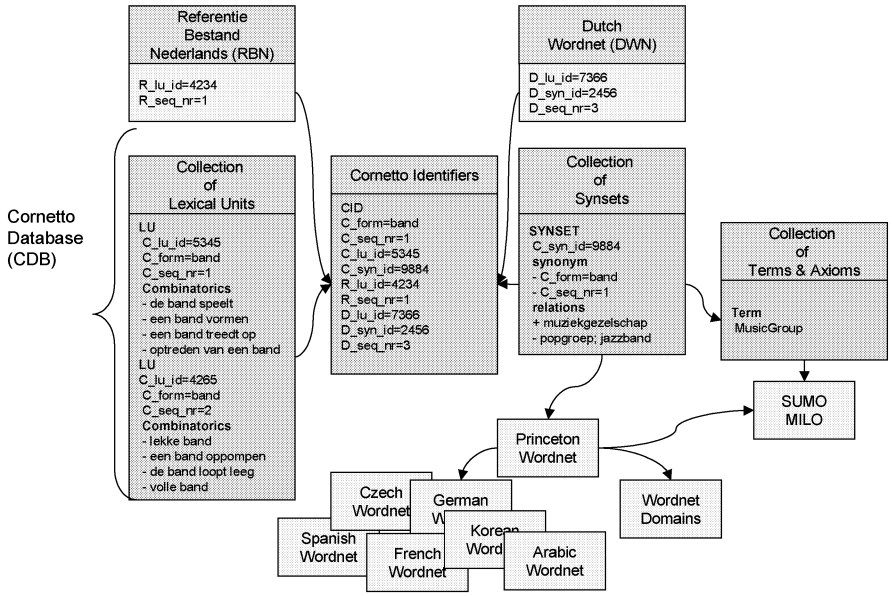
- de vorm een abstractie is van bepaalde realisaties als woord;
- de woordsoort van die realisaties gelijk is;
- de betekenis identiek is, waarbij die betekenis wordt vastgelegd door een equivalentie relatie met een formele ontologie of een KIF expressie met gebruikmaking van die ontologie;

De uiteindelijke databank zal bestaan uit een combinatie van drie datacollecties met conceptuele structuren:

1. een verzameling van Lexical Units met combinatorische informatie;
2. een verzameling Synsets met conceptuele relaties in de vorm van een wordnet;
3. een verzameling ontologietermen in de vorm van een ontologie met axioma's;

Iedere subdatabase vertegenwoordigt een eigen paradigma en heeft een eigen karakter, wat verschillende toepassingsmogelijkheden biedt. In figuur-6 wordt een schematisch overzicht gegeven van de database. De verzameling LUs zal voornamelijk worden overgenomen uit het RBN en de verzameling synsets uit het DWN.

In het schema worden ze aan elkaar gekoppeld door middel van zogenaamde Cornetto Identifiers. Deze identifiers geven precies aan welke woordbetekenissen en synsets uit de oorspronkelijke databases bij elkaar horen. In dit voorbeeld wordt dus de eerste betekenis van *band* uit het RBN, aangegeven als R_seq_nr=1, gekoppeld aan de derde betekenis uit het DWN: D_seq_nr=3. Het resultaat is een LU in de Cornetto database die de eerste betekenis vormt: C_seq_nr=1, in de betekenis van een *muziekband*. De LU bevat de combinatorische informatie voor *band*: *spelen*, *vormen*, *optreden*, etc, en de synset de conceptuele semantische informatie: een soort *muziekgezelschap* en subtypes/hyponiemen *popgroep* en *jazzband*. Als contrast wordt hier als tweede betekenis gegeven *band* om een *wiel*. Deze betekenis vormt een andere LU.



Figuur 6: Architectuur van de Cornetto database

De synset *band* is verder gekoppeld aan zowel de ontologie als aan het Engelse WordNet. Aangezien het Engelse WordNet nu nog als ILI wordt gebruikt voor de koppeling van de verschillende wordnets, blijft het Nederlandse wordnet binnen de Cornetto database dus ook verbonden met en dus verankerd aan alle andere wordnets in de wereld. De definities aan de hand van de ontologie zullen grotendeels bestaan uit SUMO met de uitbreiding MILO, en

eventueel verder aangevuld met disjuncte types in zoverre die nodig zijn om rigide concepten te onderscheiden van antirigide rollen.

4.2 De relatie tussen woorden en concepten in Cornetto

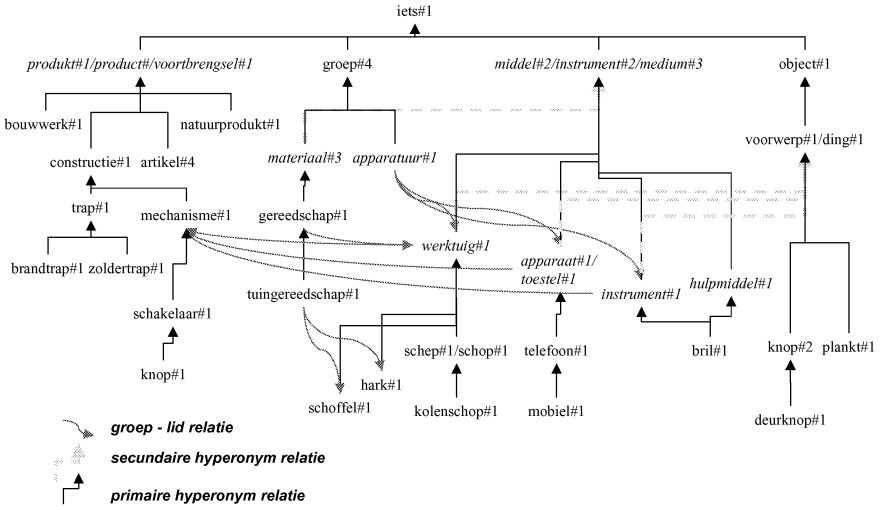
Door de expliciete scheiding van 3 soorten conceptuele databases is het mogelijk om de relaties tussen woord en concept in detail te bestuderen. Wat is een concept en wat is een woord? Is ieder woord een concept en hoe combineren woorden in een bepaalde betekenis?

De lijst van woorden in een taal ligt niet vast. Woorden die uit delen bestaan zoals samenstellingen (bijv. *zout-vaat-je*) en afleidingen (bijv. *echt-heid*) worden vaak alleen als woorden beschouwd indien de betekenis op een of andere manier niet helemaal afleidbaar is uit de delen: *kap-stok* of *be-zoek*. Dat zou echter betekenen dat woorden zoals *doodslaan* en *fijnknijpen* niet zouden hoeven worden opgenomen in het lexicon. Toch staan deze woorden in een algemeen woordenboek en mijns inziens met rede. Er spelen namelijk nog meer factoren een rol dan alleen de transparantheid van betekenis. Zo is het nog maar de vraag of mensen frequent gebruikte samenstellingen en uitdrukkingen ook daadwerkelijk 'samenstellen' uit de delen. Het is waarschijnlijker dat deze woorden en begrippen *ready-made* zijn opgeslagen in ons geheugen.

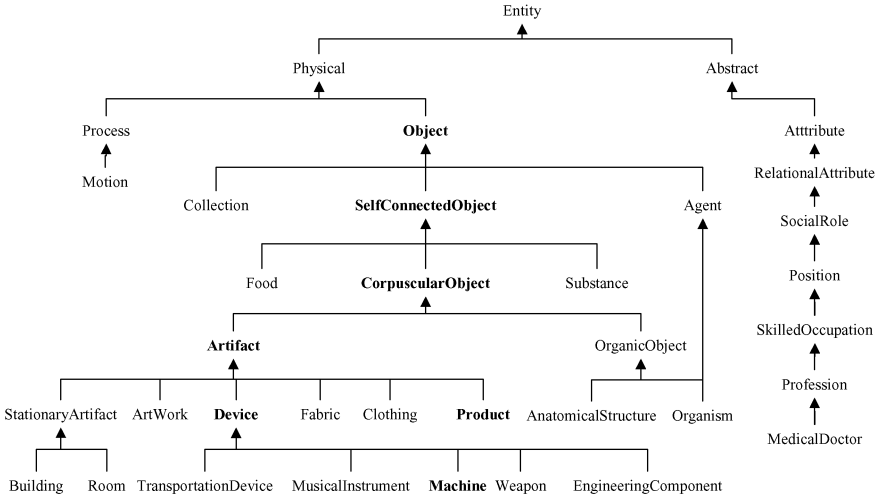
In plaats van te streven naar een 'woestijnlandschap' van primitieve begrippen, zou een wordnet gezien moeten worden als een zo rijk mogelijk netwerk met de *beschikbare* woorden en uitdrukkingen in een taal. ***Beschikbaarheid*** is daarbij een sleutelbegrip en wordt bepaald door allerlei factoren, waaronder frequentie, mode, subculturen maar ook autorisatie: publieke personen in een taalgemeenschap, zoals bijvoorbeeld politici, kunnen waarschijnlijk makkelijker nieuwe woorden introduceren. ***Beschikbaarheid*** van woorden en uitdrukkingen is een belangrijk stuk informatie in een lexicon dat door sprekers wordt gebruikt in de keuze van woorden en taalconstructies. De spreker weet min of meer welke woorden de hoorder kent. Hij anticipeert op de beschikbaarheid van het vocabulaire en dat zal de keuzes bepalen om bepaalde dingen op een bepaalde manier te zeggen. Een lexicon en zeker een wordnet zal een accurate weergave moeten zijn van het beschikbare vocabulaire gegeven een bepaald concept. Het is een verzameling aan ***non***-creativiteit en clichés om aan dingen in de werkelijkheid te refereren en ook om ze op een bepaalde manier te benaderen en te perspectiveren of te parafraseren. De structuur en de elementen van de verzameling zullen zelfs van invloed zijn op ons denken.

Dit heeft ook consequenties voor de wijze waarop we een wordnet aan een ontologie koppelen. In de volgende twee figuren worden een fragment van het Nederlandse wordnet en een fragment van de bovenlaag van SUMO getoond. Niet alle concepten worden hier weergegeven om de figuren leesbaar te houden. Als we eerst naar wordnet kijken dan zien we dat er een grove indeling wordt gemaakt in 3 clusters van concepten: *producten*, *middelen* en *voorwerpen*. Het concept product is vrij abstract. Het bevat allerlei subconcepten waar de nadruk ligt op het productieproces. De processen zelf variëren van natuurlijke processen, abstracte processen (bijv. *geestesproduct*), of industriële processen. Onder middel vinden we ook een bonte verzameling van subklassen zoals *werktuig*, *apparaat*, *instrument* en *hulpmiddel*. Daarnaast zijn er nog collectieven of verzamelingen aan instrumenten zoals *gereedschap* en *apparatuur*. Onder voorwerp vinden we vele honderden specifieke concepten die vaak niet verder zijn onderverdeeld. Verder wordt voor veel *middelen* een tweede relatie gegeven naar *voorwerp*, hier aangegeven door een licht gestippelde pijl. Het gaat hier dus om

zowel *middelen* als *voorwerpen*. Zo ook bestaan de verzamelingen instrumenten uit instrumentobjecten.



Figuur 7: Top-hiërarchie van het Nederlandse wordnet



Figuur 8: Deel van het top-niveau van de SUMO ontologie

Kijken we nu naar de SUMO top-ontologie in figuur-8 dan zien we een heel andere indeling van de werkelijkheid. In SUMO wordt eerst een indeling gemaakt in objecten (**CorpuscularObject**) en daarna in **Artifact** en vervolgens in **Product**. Hier wordt dus gesteld dat alle artefacten objecten zijn en alle producten artefacten. Deze indeling is dus heel anders dan de hiërarchie van het Nederlandse wordnet.

We zien hier aan de ene kant dat de ontologie een meer heldere en eenvoudige indeling maakt dan het Nederlandse wordnet maar aan de andere kant bevat het Nederlandse wordnet patronen en relaties die subtieler en misschien wel informatiever zijn dan de ontologie. Zo zou je op het eerste gezicht kunnen zeggen dat *werktuig*, *apparaat*, *instrument* en *hulpmiddel* net zo goed een klasse zouden kunnen vormen. Dan staan namelijk alle *Devices* bij elkaar en kan de semantiek in een keer worden uitgedrukt. Bovendien zijn veel van die *Devices* ook voorwerpen en ook artefacten dus lijkt de hiërarchische ordening van SUMO object -> artefact -> device, meer efficiënt. Toch lijkt het alsof de groepjes woorden die we eronder vinden een suggestieve indeling weergeven. *Hulpmiddel* bevat voornamelijk middelen die hulp bieden omdat iets moeilijk is, iemand iets niet goed (meer) kan, iets gecorrigeerd moet worden: *beugel*, *looprek*, *bril*. De instrumenten zijn daarentegen vaak technischer dan andere klasses: *collimator*, *conductor*, *gastroscoop*, *heliograaf*. Werktuigen zijn weer niet elektrisch: *appelboor*, *bandenlichter*, *beitel*, *bezem*, *bijl*, en apparaten/toestellen vaak weer wel: *camera, centrifuge, computer, condensator, detector, diepvriezer*, maar niet altijd: *aansteker, brandblusser, brander*. Als men de lijsten bekijkt dan zijn ze niet altijd even consequent maar toch lijkt er wel een lijn in te zitten. Het is op zich niet vreemd dat een *bril* zowel een hulpmiddel als een instrument is. Kijken we naar de *voorwerpen* dan zien we dat de woorden die daar gegeven worden (*knop* en *plank*) een sterkere vormassociatie hebben en een meer open functie in vergelijking tot de *apparaten*. Vergelijk bijvoorbeeld *knop* in de betekenis van *deurknop* of *versiersel* dat erg vorm bepaald is en *knop* in de betekenis van *schakelaar* dat naar *mechanisme* gaat en meer functie bepaald is.

De indeling is subtiel en niet direct hard te maken in de ontologie, maar dat is dan wellicht de functie die een wordnet heeft in toevoeging tot de ontologie. Waar de ontologie misschien dezelfde implicatie zou afleiden voor al deze concepten (artefact object met een bepaalde instrument rol) geeft een wordnet informatie over welke woorden met elkaar in competitie zijn binnen dezelfde subklasse. Bovendien geeft het beter aan dat we wel met *werktuig* naar *hark* kunnen verwijzen maar niet met *instrument* of *apparaat*. Een wordnet voorspelt dus beter welke woorden er ***beschikbaar*** zijn vanuit een bepaald concept om naar bepaalde dingen te verwijzen, zonder dat dat in de ontologie wordt hardgemaakt. Uit welke parafrases en rollen kun je kiezen, met welke types is een concept in competitie?

Een ander probleem heeft te maken met de ordening van de concepten in SUMO. Doordat Product en Artefact zo laag in de hiërarchie zitten worden ze onbruikbaar om toe te passen op het Nederlandse wordnet. De producten in het Nederlandse wordnet zijn niet noodzakelijkerwijs artefacten, vergelijk *natuurproduct*, en ook niet altijd objecten, vergelijk *brouwsel*:

_ product_1/product_1/voortbrengsel_1

assemblage, bouwwerk, constructie, smeedwerk, werk

artikel, fabrikaat, halffabrikaat, industrieproduct, serieproduct

afbraakproduct, afvalproduct, bijproduct, massaproduct, kwaliteitsproduct, kunst, primeur, eindproduct, exportproduct, geesteskind, gewrocht, misbaksel, specialiteit, diepvriesproduct

ei, graanproduct, genproduct, kweeksel, landbouwproduct, natuurproduct, olieproduct, teelt, tuinbouwproduct, zuivelproduct,

brouwsel, condensaat, derivaat, distillaat, mengsel, verbrandingsproduct, versproduct, vormsel,

Een ontologische definitie van product in deze abstracte betekenis is moeilijk te geven. In SUMO is immers alles het resultaat van een proces en daarom een product. Het kan dus niet worden gebruikt om een type Product te definiëren dat zich onderscheidt van het hoogste niveau: *Entity*. Formeel semantisch is er geen onderscheid. Zeggen dat iets een product is in de abstracte zin, is eigenlijk helemaal niets zeggen volgens SUMO.

Tenslotte is er in het Nederlands een probleem met de relatie tussen *Artefact* en *CorpuscularObject*. Kennelijk zijn alle *artefacten* in de Anglo-Saxische cultuur ook *voorwerpen*. In het Nederlands is er echter geen directe equivalent voor *Artefact* in die specifieke betekenis. Het woord *artefact* is veel algemener en kan zelfs naar abstracte zaken verwijzen, als in “een artefact van het menselijk denken”. Daarentegen kent het Nederlands wel een woord als *kunststof* waarvoor weer geen equivalent in het Engels bestaat. *Kunststof* kan echter niet worden ondergebracht in SUMO omdat *substanties* exclusief zijn onderscheiden van *voorwerpen*. We zien dus dat, alhoewel een ontologie taalneutraal beoogt te zijn, er toch culturele en zelfs Engelstalige invloeden zijn die de structuur bepalen en te specifiek maken.

Zonder dat we de eigen structuur van het wordnet hoeven los te laten is het toch mogelijk om de noodzakelijke ontologische labels en definities toe te voegen aan de concepten. Indien er geen directe equivalentie bestaat tussen concepten in de ontologie dan kunnen de semantische implicaties ook aan specifiekere concepten worden overgedragen. In het Nederlandse wordnet kunnen we van ieder woord apart aangeven of het behoort tot het type *CorpuscularObject* of *Artefact*. Het spreekt voor zich dat dit niet nodig is als deze implicatie al blijkt uit de semantisch structuur van wordnet zelf. Mocht het zo zijn dat iets artificieel is maar geen *Object*, dan kunnen we of de ontologie uitbreiden met het type *ArtifactSubstance* of we kunnen de woorden in het Nederlands, zoals kunststof, apart voorzien van een expressie die aangeeft dat kunststoffen het resultaat zijn van een *Making* proces:

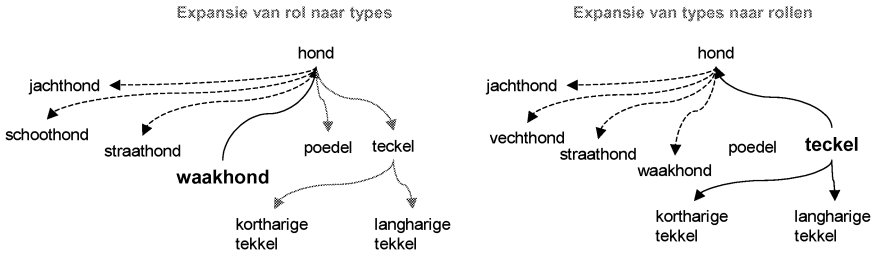
```
(<=>
   (instance ?ARTIFACT Substance)
   (exists (?MAKING)
      (and
         (instance ?MAKING Making)
         (result ?MAKING ?ARTIFACT))))
```

Dat neemt niet weg dat de discussie over de structuur van een ontologie als SUMO gevoed zal moeten worden door de concepten en relaties die andere talen dan het Engels met zich meebrengen. Alleen dan kan een dergelijke ontologie zich kandidaat stellen als een taalneutrale en universele index voor alle talen.

5 Een zinnig gesprek met een computer

Een taaltechnologie bedrijf als Irion Technologies heeft elke dag te maken met situaties waarin er met tekstsymbolen wordt geschoven, hetzij omdat klanten enorme hoeveelheden tekst moeten verwerken en rubriceren, hetzij omdat er systemen moeten worden gemaakt waarbij redelijk naïeve gebruikers hun weg moeten vinden in die brij aan informatie. Vaak gaat het daarbij ook nog om teksten in verschillende talen. Irion Technologies ontwikkelt een heel scala aan toepassingen die die informatie beter toegankelijk maken, waaronder tekstclassificatie, zoekmachines en dialoogsystemen die mensen vragen laten stellen aan het systeem dat dan de best passende informatie geeft.

De producten van Irion maken intensief gebruik van een wordnet database voor verschillende talen. Het is daarbij belangrijk dat variaties in de manier waarop mensen over dezelfde dingen praten/schrijven worden herkend. Het gaat daarbij niet alleen om synoniemen maar met name om parafraseringen. Een ontologisch zuiver wordnet is daarbij van cruciaal belang. In figuur-9, wordt bijvoorbeeld aangegeven hoe een woord in een tekst kan worden geëxpandeerd naar gerelateerde woorden in een semantisch netwerk of wordnet, mits de relaties juist worden geïnterpreteerd. In dit voorbeeld worden twee verschillende soorten expansies weergegeven: aan de linkerkant staat het woord *waakhond* dat een rol aangeeft (antirigide), en aan de rechterkant wordt uitgegaan van het woord *teckel* dat een disjunct type is (rigide):



Figuur 9: Expansie van woorden via verschillende soorten hyponymy relaties

Voor beide situaties geldt dat ieder woord in principe geparafraseerd kan worden door een algemener woord, i.e. de hyperonym of het type waar het een subklasse van is (*hond*), maar ook door zijn eigen directe subtypes (voor *teckel* zijn dat bijvoorbeeld *kortharige* en *langharige teckel*). Iemand die dus zoekt naar *teckel*, wil ook teksten vinden met *kortharige* en *langharige teckels* en eventueel teksten waar gesproken wordt over *honden*.

Als de ontologische typering echter juist is dan kunnen ook andere woorden in aanmerking komen. Omdat *waakhond* een rol is kan het vrijelijk combineren of worden toegepast op alle disjuncte typen *honden*, aangegeven door de dichte pijlen, dus ook siblings van *waakhond* als *poedel* en *teckel*. Ieder type hond kan als een waakhond optreden, wat niet wegneemt dat sommige honden beter in die rol zullen functioneren. Zoekt iemand dus naar *waakhond* dan kunnen we *poedels* en *teckels* niet uitsluiten. Andere rollen zijn misschien zwakker maar niet

onmogelijk (ze zouden lager kunnen scoren in een zoekmachine). Aan de rechterzijde zien we dat het woord *teckel* niet naar andere rigide concepten kan verwijzen, zoals *poedel*, maar wel naar alle rollen.

Het onderscheid tussen rollen en types kan dus worden gebruikt om een meer verfijnde expansie van een woord naar mogelijke andere woorden te geven. De expansie van woorden naar varianten is dus gebaat bij het gebruik maken van de koppeling tussen de ontologie en het wordnet met de woorden uit een taal zoals hierboven beschreven. We kunnen nog een stap verder gaan. De contextualisering die door rolaanduidende woorden zoals vechthond en waakhond wordt uitgedrukt kan ook gebruikt worden bij het vinden van relevante informatie. Vragen die gesteld worden naar contexten, zoals *waken*, *bewaking beveiliging*, zullen dus ook direct matchen met een woord als wachthond dat die situatie juist contextualiseert.

Taaltechnologie speelt verder op twee niveaus een rol binnen Irion:

1. Woorden en woordcombinaties in teksten worden omgezet naar concepten en concepten naar kennis en informatie;
2. Communicatieve linguïstische modellen worden gebruikt zodat mensen kennis en informatie kunnen vinden door simpelweg vragen te stellen aan een computer in natuurlijke taal.

In beide gevallen is de relatie tussen woord en concept, zoals we die hierboven besproken hebben belangrijk.

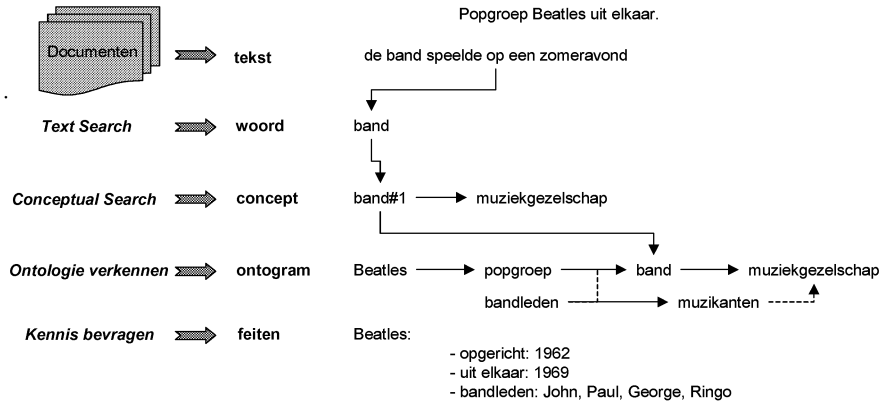
5.1 Van woord naar concept naar gestructureerde kennis

De technologie die Irion gebruikt is concept-driven. Dat wil zeggen dat kennis en informatie die in tekstuele vorm wordt aangeboden zoveel mogelijk wordt omgezet naar een conceptuele representatie die niet afhankelijk is van de bewoording die gebruikt is. Nadat woorden in teksten zijn gerelateerd aan concepten kunnen rijkere kennisstructuren en slimmere indexen worden gebouwd om informatie te ontsluiten en gebruiken.

Dit proces is schematisch weergegeven in figuur-10. De tekst in een verzameling teksten wordt verregaand geanalyseerd. Dit houdt in dat belangrijke woorden worden gedetecteerd maar ook allerlei relaties tussen die woorden, bijvoorbeeld *band* => *spelen*. Op het woordniveau is het mogelijk om te zoeken naar woorden in de teksten maar dat is niet voldoende. Op grond van de woorden in de omgeving van *band* kunnen we bepalen dat het hier om een *muziekgezelschap* gaat. Zoals we gezien hebben, kunnen we hiervoor de combinatorische informatie uit Cornetto gebruiken.

Weten we eenmaal dat het hier om een *muziek* band gaat, dan kunnen we conceptueel zoeken. We vinden deze zinnen ook als je zoekt naar *popgroep, bandleden* of *muzikanten*. De volgende stap is dan om uit de tekst een ontologische structuur te halen die weergeeft welke concepten en relaties in de tekst worden uitgedrukt. Hiervoor wordt kennis gebruikt uit wordnet en uit een ontologie. We kunnen uit de tekst afleiden dat de *Beatles* een *popgroep* zijn (Popgroep Beatles) en het semantisch netwerk leert ons dat een *popgroep* een soort *band* is en dat een *band* uit *bandleden* bestaat die weer *muzikanten* zijn. De ontologie leert ons wat *oprichten* en *uit elkaar gaan* kan betekenen voor *gezelschappen*. Het resultaat is een rijke semantische structuur die op maat gesneden is voor een verzameling teksten. Een dergelijke structuur noem ik een ***ontogram***. Een ontogram wordt afgeleid uit een generieke ontologie op

grond van de woorden die in een verzameling teksten voorkomen, eventueel aangevuld met nieuwe concepten uit de teksten. Het is mogelijk om een ontogram weer te geven als een boomstructuur waar je doorheen kunt lopen om de concepten en de relaties te bekijken, zoals in figuur-10 het op-een-na-laagste niveau. Ieder concept in een ontogram is weer verbonden met de woorden en uitdrukkingen in de tekst. Door middel van een zogenaamde deep-link, kan een gebruiker met een muisklik direct van concept doorklikken naar alle tekstvoorkomens van dat concept in alle mogelijke bewoordingen. Het systeem springt dan van een concept terug naar de index van woorden waaruit het concept in eerste instantie was afgeleid. Omdat de opgebouwde ontologie overzichtelijk en compact is, biedt het een goede manier om inzicht te krijgen in de informatie die in de teksten te vinden is.



Figuur 10: Cascade van kennis- en informatie-extractie

Tenslotte kan de rijke semantische structuur zelf gebruikt worden om dat wat er beweerd wordt in de tekst in een feitendatabank op te slaan als kennis. Zogenaamde kennisontginners (data-mining programma's) kunnen de ontologie gebruiken om bepaalde relaties in teksten te detecteren en verifiëren. Daarbij kunnen verschillende stukjes informatie aan elkaar gekoppeld worden. Zo kan het programma bijvoorbeeld uit allerlei informatie afleiden dat de tekst gepubliceerd is in 1969 en dat het om een nieuwsbericht gaat. Indien het in de teksten verder geen aanwijzingen vindt wanneer de Beatles uit elkaar zijn gegaan, dan kan het het jaartal van publicatie als suggestie opslaan (bijvoorbeeld met een lage betrouwbaarheidsscore).

De hierboven geschetste analyse lijkt heel specialistisch en gedetailleerd en daardoor foutgevoelig. Het is echter belangrijk om te realiseren dat het meestal wordt toegepast op grote hoeveelheden data en dat alleen informatie die betrouwbaar genoeg is of uit meerdere bronnen bevestigd kan worden hoeft te worden gebruikt. Dit soort extractie wordt op dit moment bijvoorbeeld toegepast voor een aantal klanten bij Irion die zeer specialistische databases aanleggen over miljoenen bedrijven en hun producten, wereldwijd. Uit de websites van die bedrijven worden automatisch de producten geëxtraheerd maar ook bijvoorbeeld de adresgegevens van de bedrijven. De producten worden in de vorm van een hiërarchie of productclassificatie opgeleverd zodat bedrijven met dezelfde producten kunnen worden

gegroepeerd en vergeleken. Verder wordt nieuws- en productinformatie verzameld en gekoppeld aan de productclassificatie.

5.2 Conceptuele informatie in een communicatief systeem

De trapsgewijze extractie van informatie en kennis uit teksten biedt dus ook de mogelijkheid om met steeds geavanceerdere technieken toegang te krijgen tot de kennis en informatie: zoeken op tekst, zoeken naar concepten, verkennen van een ontogram, en het exploreren van een feitendatabank. Een van de problemen waar we bij Irion echter tegen aanlopen is dat de mensen die die kennis willen opvragen of exploreren niet in staat zijn of niet de moeite willen nemen om de rijke en soms complexe structuren te begrijpen en te leren kennen. Dit beperkt het gebruik tot mensen binnen een organisatie en sluit het grotere publiek uit. Bovendien kan het zo zijn dat de woorden in de tekst niet aansluiten bij de woorden van de gebruiker. Zo zien we bijvoorbeeld bij juridische teksten en regelgeving van de overheid dat vaak algemene bewoordingen worden gekozen, terwijl een gebruiker zijn eigen specifieke geval of case beschrijft:

Regelgeving:	***Voertuigen*** op de ***openbare weg*** moeten worden voorzien van een geldig ***kentekenbewijs*** en duidelijke ***markeringen***;
Gebruikersvraag:	Ik heb een ***oldtimer*** die ik nooit gebruik. Moet die ook een ***kentekenplaat*** en ***lichten*** hebben als ik die op ***straat*** parkeer?

Je kunt de informatie over de regelgeving alleen vinden indien het systeem *oldtimer* matcht met *voertuig*, *openbare weg* met *straat*, *kentekenbewijs* met *kentekenplaat* en *markeringen* met *lichten*.

Naast een verschil in woordkeuze en algemeenheid kan ook het perspectief van de informatieaanbieder verschillen van het perspectief van de informatiezoeker. Vergelijk het met een website waar een schat aan informatie staat. Deze informatie is ingedeeld in webpagina's waarbij iedere pagina weer verder ingedeeld kan worden in subpagina's. Je krijgt dan een soort kaart met de vertakkingen van een boom. Als voorbeeld nemen we de website van de Letteren Faculteit van de VU, waarvan hieronder een stukje van de sitemap wordt weergegeven. De sitemap is een hiërarchische structuur die op een bepaalde manier is ingedeeld. Zo zie je dat hier gekozen is om eerst een indeling te maken in Bachelor- en Masteropleidingen en daarbinnen op inhoudelijke gronden, bijvoorbeeld de opleiding Nederlands en Literatuurwetenschap. Binnen dat niveau zie je vervolgens weer indelingen die deels op elkaar lijken (Beroepsperspectieven) en deels afwijken (Excursie, Frame). In dit geval wordt de gebruiker gedwongen om deze indeling te volgen om bij de betreffende informatie te komen. Iemand die bijvoorbeeld geïnteresseerd is in de Beroepsperspectieven van alle opleidingen zal dus een voor een de takken van de boom moeten aflopen voor iedere afdeling naar de specifieke Beroepsperspectieven van iedere opleiding: 14 Bachelor-opleidingen en 20 Masters.

Bacheloropleidingen	Masteropleidingen
Bacheloropleiding Nederlands	Masteropleiding Nederlands
Bachelorprogramma	**Beroepsperspectieven**
Beroepsperspectieven	In deeltijd
Excursie	Introductie
In deeltijd	Kosten
Introductie	Masterprogramma
Kosten	Medewerkers
Medewerkers	Meer informatie
Meer informatie	Onderzoek
Onderzoek	Toelating en inschrijving
Toelating en inschrijving	
Bacheloropleiding Literatuur	Masteropleiding Literatuur
Bachelorprogramma	Beroepsperspectieven
Beroepsperspectieven	Contact
Contact	In deeltijd
Frame	Interessante links
In deeltijd	Introductie
Interessante links	Kosten
Introductie	Masterprogramma
Kosten	Medewerkers
Medewerkers	Meer informatie
Meer informatie	Onderzoek
Tijdschrift frame	Tijdschrift frame
Toelating en inschrijving	Toelating en inschrijving
Vakkenoverzicht	

Om dit soort problemen te voorkomen hebben we bij Irion een systeem ontwikkeld dat iemand naar de juiste informatie kan leiden ongeacht de woorden die iemand kiest en de wijze waarop die informatie is ingedeeld. Dit systeem maakt gebruik van een dialoog in natuurlijke taal. De gebruiker stelt een vraag in zijn eigen woorden en vanuit zijn eigen perspectief en het systeem vergelijkt dat met de beschikbare informatie en geeft suggesties aan de gebruiker of stelt vragen ter verduidelijking.

Dit dialoogsysteem hanteert een communicatiefmodel als basis waarin 4 verschillende informatielagen worden bijgehouden:

1. De **intentie** van de gebruiker
2. De mate waarin iemand tevreden is (**satisfaction rate**).
3. De **emotionele staat** van de gebruiker; is iemand boos, vrolijk, vriendelijk.
4. De **informatie staat** gebaseerd op de inhoudelijke beschrijving die een gebruiker geeft van een informatiebehoefte; Waar is iemand naar op zoek?

Het systeem maakt een linguïstische analyse van de uitingen van een gebruiker. Na iedere analyse zal het systeem een update maken van de 4 bovengenoemde lagen. De globale strategie van het systeem is te achterhalen met welke bedoeling of intentie een gebruiker op zoek is naar informatie of een dienst. Daarbij kijkt het systeem voortdurend naar de emotionele staat en de tevredenheid van de gebruiker. Raakt de gebruiker gefrustreerd of

verloopt de communicatie moeizaam omdat de aangeboden informatie steeds niet voldoet, dan zal het systeem dat gebruiken bij het maken van de volgende keuzes.

Bij het vergelijken van de informatie die de gebruiker aandraagt en de informatie die het systeem in de aangeboden informatie heeft, wordt er voortdurend gekeken naar hoe goed de vraag op het antwoord past. Vraagt iemand in het voorbeeld van de VU website naar "beroepsperspectieven" dan zal het systeem zien dat die categorie op twee plaatsen voorkomt. Er is dus geen eenduidigheid en het systeem zal vragen om verduidelijking of meerdere opties aanbieden: *bedoel je beroepsperspectieven voor Nederlands of Literatuur?* Bij iedere vraag wordt een afweging gemaakt in hoeverre die vraag vaag is of ambigue of dat de aangeboden informatie vaag is of ambigue.

Hetzelfde gebeurt ook op het concept-woord niveau. Indien iemand als vraag stelt: "Wat voor banen kan ik krijgen met literatuurwetenschap?", dan zal *beroepsperspectief* moeten worden gerelateerd aan *banen*. Ook hiervoor gebruikt het systeem het Nederlandse wordnet. Zodra die koppeling gemaakt is, is de vraag eenduidig en kan de vragensteller meteen naar de relevante sectie worden doorverwezen, zonder dat iemand door de vooropgestelde hiërarchie moet bladeren.

Het dialoogsysteem kan dus gebruik maken van de hiërarchische indeling door de gebruiker keuzes te geven uit die indeling, maar het kan ook iemand direct naar het doel brengen ongeacht die indeling en volgens het perspectief van de gebruiker (door niveaus over te slaan).

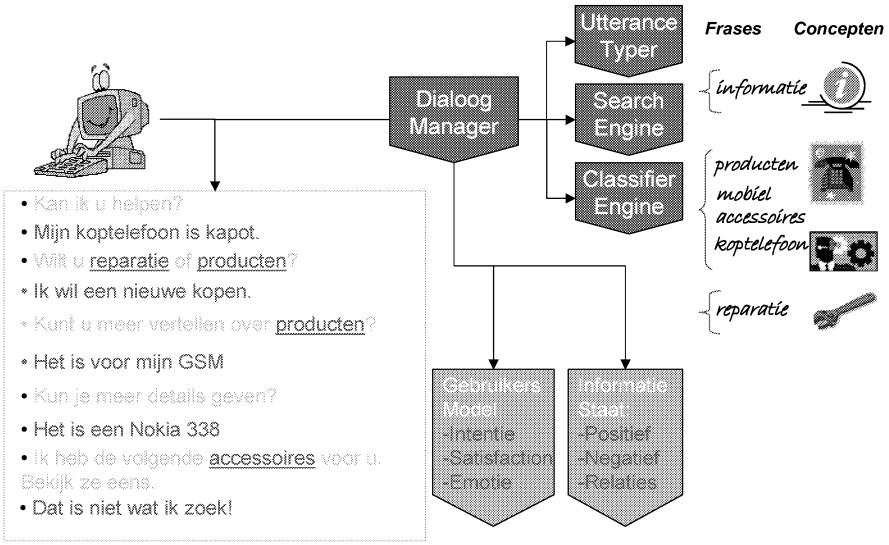
De tevredenheid van de gebruiker en de emotionele signalen worden gebruikt om te zien of het systeem wel op de goede weg zit. De dialoog lijkt daarin een beetje op het spelletje "Ik zoek, ik zoek wat jij niet ziet en het is ...", waarbij het systeem suggesties geeft op tips van de gebruiker, en de gebruiker met warm of koud antwoordt als een suggestie meer of minder in de richting gaat. Warm betekent dat de gebruiker keuzes bevestigt en tevreden is of blijft, en koud indien de gebruiker suggesties ontkent en/of boos of ontevreden wordt.

In het onderstaande plaatje wordt een hypothetisch voorbeeld gegeven van een dialoog waarin deze aspecten naar voren komen. Het systeem is zo ingericht dat de aangeboden informatie uit teksten is geanalyseerd en herkend als een aanbod aan producten met daarbij de verschillende services die worden aangeboden: *informatie*, *producten*, *accessoires* of *reparatie*. Deze informatie is deels automatisch uit de teksten afgeleid en deels gebaseerd op wat de leverancier belangrijk vindt. In dit voorbeeld gaat het om een website waarin elektronische apparaten, zoals mobiele telefoons, computers, mp3-spelers, etc. worden aangeboden.

De strategie van het systeem is om naar aanleiding van de eerste mededeling te bepalen wat de mogelijke intentie is. Op grond van "is kapot" kan het systeem bepalen dat iemand eventueel op zoek is naar ***reparatie*** als een dienst. Dit gebeurt door een automatische classificatie van de vraag/mededeling ingedeeld naar de verschillende intenties. Tegelijkertijd zal de classificatie van "koptelefoon" ook de suggestie geven dat iemand eventueel geïnteresseerd is in het kopen van een product. Aangezien het systeem niet kan kiezen tussen beide intenties, legt het de gebruiker de keuzemogelijkheid voor.

De gebruiker geeft geen direct antwoord op de keuze. Het systeem ziet dat er geen sprake is van een ontkenning en probeert met de toegevoegde informatie een betere keuze te maken. In dit geval is het zeker dat "een nieuwe kopen" duidt op het kopen van een product. Het

beschouwt de intentie als zijnde bepaald en gaat nu proberen om binnen het aanbod van producten te zoeken naar een eenduidig resultaat. Aangezien een ***koptelefoon*** bij meerdere producten kan horen vraagt het systeem om meer informatie.



Figuur 11: Globale architectuur dialoogsysteem met communicatief model

Het systeem blijft daarmee doorgaan totdat het denkt dat het zeker genoeg weet over wat er gevraagd wordt. Hierbij wordt al gebruik gemaakt van veel kennis die is opgebouwd uit teksten die zijn geïndexeerd en met gebruikmaking van de ontologie en wordnet. Zo moet het systeem weten wat de *producten* zijn maar ook dat producten *accessoires* kunnen hebben, d.w.z. producten die bij een ander specifiek product horen. Verder moet het bepalen wat de betekenis is van *mobiel* zodat het weet dat het om een telefoon gaat en dat het woord *GSM* uit de vraag kan worden gerelateerd aan *mobiel* in het aanbod. Verder zal het ook merknamen en productseries zoals *Nokia 338* moeten kunnen relateren aan gevonden concepten zoals *mobiel* en *GSM*. Uiteindelijk moet het systeem nog afleiden dat een *koptelefoon* een *accessoire* is van een *mobiel*.

Om tenslotte het plaatje af te maken: uit de laatste mededeling zal het systeem afleiden dat de gebruiker toch niet tevreden is en vervolgens terugvallen op eerdere keuzes en alternatieven aanbieden. Bij het analyseren van de uitingen van de gebruiker wordt gebruikt gemaakt van gedetailleerde lexica en grammatica om de verschillende informatiesoorten te herkennen. Zo is het bijvoorbeeld erg belangrijk om negatie te detecteren en ook de scope van de negatie: ik wil een *koptelefoon* en **geen** *mobiele telefoon*. Informatie binnen een negatieve scope wordt bijvoorbeeld weer gebruikt om te bepalen wat de gebruiker niet wil. Het resulteert in een negatieve informatiestaat.

Klassieke dialoogmodellen, zoals bijvoorbeeld systemen om vliegtickets te boeken of hotelkamers te reserveren, maken altijd gebruik van een van tevoren bepaald model van de wereld. De soort vragen en antwoorden liggen daarbij vast. Kenmerkend voor het dialoogsysteem van Irion is dat er niet uitgegaan wordt van een vastgelegd wereldmodel. Dit wil overigens niet zeggen dat de aangeboden informatie niet gestructureerd wordt, zoals we hebben gezien. Het wil alleen zeggen dat het niet beperkt is tot een bepaald type informatie of diensten. In principe staat het systeem los van de soort informatie en diensten die worden aangeboden. Er zijn natuurlijk wel beperkingen in de kennis en intelligentie van het systeem. Wat er wel gebeurt is dat de informatie zelf tot op zekere hoogte geanalyseerd wordt en gestructureerd op een open conceptniveau. Deze structurering is mogelijk door gebruik te maken van het algemeen semantisch netwerk of wordnet en een ontologie.

Het wordnet speelt nog een andere belangrijke rol in het dialoogsysteem. Een groot probleem voor klassiek dialoogmodellen is dat het systeem moet weten wat het weet maar ook wat het niet weet. Indien iemand vraagt of die ook een hotelkamer kan reserveren dan moet het systeem weten dat objecten van het type hotelkamer niet geleverd worden. Dit soort vragen staan bekend als ***out-of-domain*** vragen.

Zonder een wordnet is het echter erg moeilijk om te bepalen of een hotelkamer niet semantisch gerelateerd is aan de objecten die wel aanwezig zijn. Een model dat alleen gebaseerd is op de gemodelleerde informatie kan geen goed oordeel vellen over informatie die daar buiten valt. Het systeem kan dan maar beperkt reageren: *ik weet niet wat een hotelkamer is*. Het systeem ***weet*** namelijk ook echt niet wat het is. Het zou dus heel goed een voor hem onbekend soort *mobiele telefoon* kunnen zijn die toevallig niet genoemd wordt in de databank.

Indien het de beschikking heeft over een wordnet dan kan het systeem ten eerste detecteren dat een *hotelkamer* een heel ander ding is maar bovendien ook met zekerheid zeggen dat dat soort producten niet geleverd worden. Het kan dus antwoorden: *wij leveren geen hotelkamers maar wel elektronische apparaten*.

Het redeneren over het semantisch netwerk gaat nog een stap verder. Stel dat uit de informatie blijkt dat er wel *mobiele telefoons* zijn maar geen *portofoons*. Iemand die dan vraagt om een *portofoon* hoeft dan niet te worden afgescheept met een out-of-domain antwoord: *wij hebben geen portofoons maar wel elektronische apparaten*, maar het systeem kan uit het algemene wordnet afleiden dat *portofoons* tot dezelfde klasse behoren als *mobiele telefoons*. Een beter antwoord van het systeem is dan bijvoorbeeld: *Nee, we hebben geen portofoons maar wel andere elektronische apparaten zoals mobiele telefoons*.

We zien dus dat de rol van woorden binnen een dergelijk systeem functioneel bepaald is. De woorden worden op grond van een betekenisdefinitie verankerd in het gedrag van het systeem.

6 Conclusie

Ik heb laten zien dat de betekenis van woorden niet hoeft te zweven in een filosofisch dilemma. Woordbetekenissen zijn niet alleen maar op drift. Ze kunnen op veel manieren worden verankerd: aan cognitieve modellen, aan een ontologie, aan elkaar, aan woorden in andere talen en aan een systeem dat een bepaalde functie heeft.

Er is nog veel werk te doen om voor alle woorden en alle soorten woorden een goede definitie te geven zodat die ook bruikbaar is voor computersystemen. Ik heb echter laten zien dat er in

internationaal verband veel werk gedeeld kan worden. Het is dan ook met name een kwestie van organisatie en standaardisatie. Verder heb ik laten zien dat we in Nederland met het Cornettoproject een grote stap zetten in de richting van een database waarin woorden veel explicieter en formeler worden gedefinieerd, zonder dat we de volledige omvang en diversiteit van taal uit het oog verliezen. Binnen Irion wordt deze informatie nu al gebruikt in een systeem dat probeert betekenis van woorden zoveel mogelijk serieus te nemen en te gebruiken voor effectieve communicatie met mensen. Degelijke systemen, hoe basaal dan ook, bieden een interessant werk- en onderzoeksterrein voor de taalwetenschap.

7 Afsluiting

Volgens sommige filosofen zijn mensen 4-D objecten. Je bent nu niet de persoon die je gisteren was en niet de persoon die je morgen zult zijn. Alles en iedereen is een proces, een pad in de tijd. Dat grillige pad heeft me langs academische kliffen en afgronden geleid, tot het bedrijfsleven aan toe. Het is een proces van veel vallen en steeds weer opstaan en nooit weten of na de volgende kam de top in zicht is. Als ik het zelf niet meer zag zitten dan waren er altijd wel mensen die me de weg wezen. Als ik worstelde en ploeterde met ideeën en problemen, dan was er altijd wel iemand die je deed inzien dat het allemaal ook niet zo belangrijk was en dat er meer tussen hemel en aarde is. Hier sta ik dan op een plek in de tijd. Ik ben het resultaat van een proces waar vele mensen invloed op hebben gehad en waar ik vele mensen dank ben verschuldigd.

Van hen die ik graag zou willen bedanken komt Willy Martin op de eerste plaats. Hij heeft ervoor gezorgd dat ik sta waar ik nu sta. In het verleden hebben onze paden elkaar al vaak gekruist. Toen ik nog aan de Universiteit van Amsterdam werkte en hij hier de vakgroep Lexicologie leidde, hadden we een gedeelde interesse, het lexicon, maar met name een intellectuele match en een gedeeld zuidelijk gevoel. We hebben in het verleden vaker plannen gesmeed om iets samen te doen. Die gelegenheid heeft zich echter nooit voorgedaan, mede door mijn vlucht naar het bedrijfsleven. Totdat Willy mij gevraagd heeft om hier op de VU zijn stokje over te nemen. Samen hebben we het Cornetto project opgestart waarin onze eerdere plannen nu toch gestalte hebben gekregen. Willy, bedankt voor je inzet om hier bij de VU voor mij dit mogelijk te maken. Ik hoop dat ik je erfgoed in de toekomst in het juiste frame kan laten doorleven.

Natuurlijk ben ik ook het College van Bestuur van de VU, de Stichting Het Vrije Universiteitsfonds en de Faculteit der Letteren veel dank verschuldigd. Ze hebben in deze dagen van reorganisatie en bezuiniging, waar alles alleen maar minder lijkt te worden, toch de moed en durf gehad om deze leerstoel Computationele Lexicologie in te voeren. Ik wil ze graag bedanken voor het gestelde vertrouwen en ik hoop dat ik samen met hen in de komende tijden constructief aan de slag kan. Het was hun idee om een brugfunctie te maken tussen de universiteit en bedrijfsleven. Om te laten zien dat voor letterenstudenten een ongekend perspectief gloort als linguïstische ingenieurs van een digitale kennis- en communicatiemaatschappij. Inmiddels werkt Willy Martin's laatste student bij Irion en ik hoop dat ik ook met deze rede een tipje van die sluier heb kunnen lichten.

Een brug naar het bedrijfsleven leunt aan een kant ook stevig op dat bedrijfsleven. Ik ben Irion Technologies dan ook veel dank verschuldigd dat ze mij de ruimte geven voor deze stap. Ik weet zeker dat ook Irion hiervan zal profiteren en dat beide partijen er sterker en beter uit zullen komen. De vele ideeën en toepassingen die bij Irion aan de orde van de dag zijn, maar waar nooit tijd voor is, kunnen hier aan de VU rijpen en onderzocht worden. Omgekeerd biedt

Irion voor de Letterenstudenten aan de VU een inspirerend perspectief om de geleerde kennis toe te passen en te toetsen aan de dagelijkse praktijk.

Nu ga ik verder terug op het pad van de tijd waarlangs ik gekomen ben. Vele collega's in binnen- en buitenland waar ik in veel projecten heb samengewerkt zijn de voedingsbodem geweest van veel van wat ik denk en weet: *Links, Acquilex I & II, Like, Sift, EuroWordNet I & II, Euroterm, Balkanet, Meaning, Arabic Wordnet, Cornetto*. Het zijn er teveel om op te noemen en ik wil ze dan ook allemaal uit de grond van mijn hart bedanken voor de fascinerende discussies en samenwerking. Ik hoop dat ik dat nog lang kan voortzetten en dat er nog vele projecten zullen volgen.

Echte diepgang bereik je misschien alleen in je proefschrift. Ik heb nog steeds het gevoel dat daar de intellectuele kern en basis ligt voor wat ik nu doe en ga doen. Simon Dik en Lachlan Mackenzie hebben mij daar beide op hun eigen wijze in gesteund. Ik ben hen daar allebei erg dankbaar voor.

Aan de wieg van dit alles staan natuurlijk degene die het meest nabij zijn in je leven. Mijn vader had altijd een groot hart voor de wetenschap en is met zijn klassieke geest altijd een groot inspirator geweest. Mijn schoonvader heeft daar op zijn eigen wijze aan bijgedragen. Als ik mijzelf weer eens verloor in onmogelijke gedachtenkronkels, kon zijn directe Amsterdamse humor mij vaak weer met beide voeten op de grond doen belanden. Ik mis hen beide.

Als allerlaatste is er de ereplaats voor Anja, Sam en Niqee. Jullie hebben het meest te lijden gehad van de vele en vaak nachtelijke werkuren en dat uitgeputte hoofd aan de ontbijttafel, of erger nog, niet aan de ontbijttafel. De vele reizen naar het buitenland, de vele deadlines, crashende computers, een congres in Korea dat ten prooi valt aan spam-filters, ze staan allemaal als horden in ons leven gegrift waar ik zonder jullie niet over was gekomen. Jullie hebben mij altijd gesteund en me laten zien waarvoor ik het deed. Dank jullie alle drie voor het hordenlopen en de ruimte die jullie me hebben gegeven. Ik hoop dat de blues het waard was.

Ik heb gezegd.

8 Referenties

Fellbaum, C. (1998, ed.) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Geerst, G. en H. Heestermans (1992) *Groot Woordenboek der Nederlandse Taal*, Van Dale Lexicografie, Utrecht.

Guarino, N. and C. Welty (2002) Identity and subsumption. In: R. Green, C. Bean and S. Myaeng (eds.), *The Semantics of Relationships: an Interdisciplinary Perspective*. Kluwer .

Regier, T., P. Kay, P. and R.S. Cook (2005) Focal colors are universal after all. *Proceedings of the National Academy of Sciences* 102:8386-8391.

Maks, I., W. Martin, and H. de Meerseman (1999) *RBN Manual*, Vrije Universiteit Amsterdam.

Niles, I., and Pease, A. (2001) Towards a Standard Upper Ontology. In: *Proceedings of FOIS* 2001, Ogunquit, Maine, pp. 2-9.

Õim H., K Vider, L. Paldre, H. Orav, K. Pala (1999) *Specification of Estonian and Czech Wordnets*, Deliverable 2D003 EuroWordNet, LE4-8283

Ogden, C.K. and I.A. Richards (1923) *The Meaning of Meaning. A Study in the Influence of Language upon Thought and The Science of Symbolism* London 1923, 10th edition 1969

Pease, A. (2000) *Standard Upper Ontology Knowledge Interchange Format*. Web document http://suo.ieee.org/suo-kif.html.

Quine, W. V. O. (1964) *Word and Object*. MIT Pres.

Searl, J. (1990) Is the Brain's Mind a Computer Program?, *Scientific American,* 262(1):26-31

Sowa, J.F. (1999) *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks Cole Publishing Co., Pacific Grove, CA.

Turing, A. (1950) Computing Machinery and Intelligence, *Mind,* LIX:433-460.

Vossen, P (1998, ed.) *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer, Dordrecht.

Vossen, P., W. Peters, J. Gonzalo (1999) Towards a Universal Index of Meaning, in: *Proceedings of ACL-99 Workshop*, Siglex-99, Standardizing Lexical Resources, p 81- 90, June 21-22, 1999, University of Maryland, College Park, Maryland, USA.

Vossen P. (2004) EuroWordNet: a multilingual database of autonomous and language-specific wordnets connected via an Inter-Lingual-Index. *International Journal of Lexicography*, Vol.17 No. 2, OUP, 161-173.

---

[i] Andere betekenissen van *doorslaan* zijn misschien meer prominent maar de betekenis van doorgaan met slaan is ook mogelijk.

[ii] Wel is het zo dat antirigide concepten en met name de zogenaamde rollen, zich meestal hoger in de hiërarchie bevinden. Als het om specifieke concepten gaat dan treffen we meer en meer rigide concepten aan. Toch zal dit met de hand geverifieerd moeten worden.

[iii] In het Limburgs is koffie maken overigens wel een mogelijke combinatie.

[iv] Een computerprogramma dat de betekenis van woorden in teksten moet bepalen zou deze informatie kunnen gebruiken door te zoeken naar deze woorden in de buurt van het woord *band.*